FORM PTO-1390 (Modified) (REV 11-98) U.S. DEPARTMENT OF COMMERCE PATENT AND TRADEMARK OFFICE

TRANSMITTAL LETTER TO THE UNITED STATES DESIGNATED/ELECTED OFFICE (DO/EO/US) CONCERNING A FILING UNDER 35 U.S.C. 371

ATTORNEY'S DOCKET NUMBER	U.S. APPLICATION NO. (IF KNOWN, SEE 37 CFR)
87977.029101	09/763149

INTERNATIONAL APPLICATION NO.	INTERNATIONAL FILING DATE	PRIORITY DATE CLAIMED
PCT/EP99/06081	19 August 1999	19 August 1998

TITLE OF INVENTION

METHOD AND DEVICE FOR THE CONCATENATION OF AUDIOSEGMENTS, TAKING INTO ACCOUNT COARTICULATION

APPLICANT(S) FOR DO/EO/US

Christoph Buskies

Applicant herewith submits to the United States Designated/Elected Office (DO/EO/US) the following items and other information:

1. ☒ This is a **FIRST** submission of items concerning a filing under 35 U.S.C. 371.
2. ☐ This is a **SECOND** or **SUBSEQUENT** submission of items concerning a filing under 35 U.S.C. 371.
3. ☐ This is an express request to begin national examination procedures (35 U.S.C. 371(f)) at any time rather than delay examination until the expiration of the applicable time limit set in 35 U.S.C. 371(b) and PCT Articles 22 and 39(1).
4. ☒ A proper Demand for International Preliminary Examination was made by the 19th month from the earliest claimed priority date.
5. ☒ A copy of the International Application as filed (35 U.S.C. 371 (c) (2))
   a. ☒ is transmitted herewith (required only if not transmitted by the International Bureau).
   b. ☐ has been transmitted by the International Bureau.
   c. ☐ is not required, as the application was filed in the United States Receiving Office (RO/US).
6. ☒ A translation of the International Application into English (35 U.S.C. 371(c)(2)).
7. ☐ A copy of the International Search Report (PCT/ISA/210).
8. ☐ Amendments to the claims of the International Application under PCT Article 19 (35 U.S.C. 371 (c)(3))
   a. ☐ are transmitted herewith (required only if not transmitted by the International Bureau).
   b. ☐ have been transmitted by the International Bureau.
   c. ☐ have not been made; however, the time limit for making such amendments has NOT expired.
   d. ☐ have not been made and will not be made.
9. ☐ A translation of the amendments to the claims under PCT Article 19 (35 U.S.C. 371(c)(3)).
10. ☒ An oath or declaration of the inventor(s) (35 U.S.C. 371 (c)(4)).
11. ☐ A copy of the International Preliminary Examination Report (PCT/IPEA/409).
12. ☐ A translation of the annexes to the International Preliminary Examination Report under PCT Article 36 (35 U.S.C. 371 (c)(5)).

Items 13 to 20 below concern document(s) or information included:

13. ☐ An Information Disclosure Statement under 37 CFR 1.97 and 1.98.
14. ☐ An assignment document for recording. A separate cover sheet in compliance with 37 CFR 3.28 and 3.31 is included.
15. ☒ A **FIRST** preliminary amendment.
16. ☐ A **SECOND** or **SUBSEQUENT** preliminary amendment.
17. ☐ A substitute specification.
18. ☐ A change of power of attorney and/or address letter.
19. ☒ Certificate of Mailing by Express Mail
20. ☒ Other items or information:

Small Entity Certification
English Translation of the International Application Claims as Amended Under Article 34 PCT

U.S. APPLICATION NO. (IF KNOWN, SEE 37 CFR)	INTERNATIONAL APPLICATION NO.	ATTORNEY'S DOCKET NUMBER
09763149	PCT/EP99/06081	87977.029101

21. The following fees are submitted:.

BASIC NATIONAL FEE (37 CFR 1.492 (a) (1) - (5)) :

		CALCULATIONS PTO USE ONLY
☐	Neither international preliminary examination fee (37 CFR 1.482) nor international search fee (37 CFR 1.445(a)(2) paid to USPTO and International Search Report not prepared by the EPO or JPO	**$1,000.00**
☒	International preliminary examination fee (37 CFR 1.482) not paid to USPTO but Internation Search Report prepared by the EPO or JPO	**$860.00**
☐	International preliminary examination fee (37 CFR 1.482) not paid to USPTO but international search fee (37 CFR 1.445(a)(2)) paid to USPTO	**$710.00**
☐	International preliminary examination fee paid to USPTO (37 CFR 1.482) but all claims did not satisfy provisions of PCT Article 33(1)-(4)	**$690.00**
☐	International preliminary examination fee paid to USPTO (37 CFR 1.482) and all claims satisfied provisions of PCT Article 33(1)-(4)	**$100.00**

ENTER APPROPRIATE BASIC FEE AMOUNT =	**$860.00**	
Surcharge of **$130.00** for furnishing the oath or declaration later than ☐ 20 ☒ 30 months from the earliest claimed priority date (37 CFR 1.492 (e)).	**$130.00**	

CLAIMS	NUMBER FILED	NUMBER EXTRA	RATE		
Total claims	66 - 20 =	46	x $18.00	**$828.00**	
Independent claims	4 - 3 =	1	x $80.00	**$80.00**	
Multiple Dependent Claims **(check if applicable)**.			☐	**$0.00**	
TOTAL OF ABOVE CALCULATIONS =				**$1,898.00**	
Reduction of 1/2 for filing by small entity, if applicable. Verified Small Entity Statement must also be filed (Note 37 CFR 1.9, 1.27, 1.28) **(check if applicable)**.			☒	**$949.00**	
SUBTOTAL =				**$949.00**	
Processing fee of **$130.00** for furnishing the English translation later than ☐ 20 ☐ 30 months from the earliest claimed priority date (37 CFR 1.492 (f)). +				**$0.00**	
TOTAL NATIONAL FEE =				**$949.00**	
Fee for recording the enclosed assignment (37 CFR 1.21(h)). The assignment must be accompanied by an appropriate cover sheet (37 CFR 3.28, 3.31) **(check if applicable).**			☐	**$0.00**	
TOTAL FEES ENCLOSED =				**$949.00**	
				Amount to be: refunded	$
				charged	$

☐ A check in the amount of to cover the above fees is enclosed.

☒ Please charge my Deposit Account No. **10-0223** in the amount of **$949.00** to cover the above fees.
A duplicate copy of this sheet is enclosed.

☒ The Commissioner is hereby authorized to charge any fees which may be required, or credit any overpayment to Deposit Account No. **10-0223** A duplicate copy of this sheet is enclosed.

NOTE: Where an appropriate time limit under 37 CFR 1.494 or 1.495 has not been met, a petition to revive (37 CFR 1.137(a) or (b)) must be filed and granted to restore the application to pending status.

SEND ALL CORRESPONDENCE TO:

Thomas R. FitzGerald, Esq.
Reg. No. 26,730
JAECKLE FLEISCHMANN & MUGEL, LLP
39 State Street
Rochester, New York 14614-1310
Tel: (716) 262-3640
Fax: (716) 262-4133

SIGNATURE

Thomas R. FitzGerald
NAME

Reg. No. 26,730
REGISTRATION NUMBER

February 16, 2001
DATE

PATENT
87977.029101

IN THE UNITED STATES PATENT & TRADEMARK OFFICE

Applicant:	Christoph Buskies	)	Examiner: Unknown
PCT International Application No. :	PCT/EP99/06081	)	
International Filing Date:	19 August 1999	)	Art Unit: Unknown
For:	METHOD AND DEVICE FOR THE CONCATENATION OF AUDIOSEGMENTS, TAKING INTO ACCOUNT COARTICULATION	)	

EXPRESS MAIL CERTIFICATE

Assistant Commissioner for Patents and Trademarks
Washington, D.C. 20231
BOX: PCT

Dear Sir:

Certificate is attached to the Transmittal Letter to the U.S. Designated Elected Office Concerning a Filing Under 35 U.S.C. § 371 (In Duplicate) (4 pgs) of the above-named application.

"EXPRESS MAIL" Label Number EL692934382US

DATE OF DEPOSIT February 16, 2001

I hereby certify that this paper or fee is being deposited with the United States Postal Service "Express Mail Post Office to Addressee" service under 37 C.F.R. 1.10 on the date indicated above and is addressed to the Assistant Commissioner of Patents and Trademarks, Washington, D.C., 20231, Box: PCT.

Penny R. Turrell
(Typed or printed name of person mailing paper or fee)

(Signature of person mailing paper or fee)

116312

PATENT
87977.029101

IN THE UNITED STATES PATENT & TRADEMARK OFFICE

Applicant:	Christoph Buskies	Examiner: Unknown
PCT International Application No. :	PCT/EP99/06081	
International Filing Date:	19 August 1999	Art Unit: Unknown
For:	METHOD AND DEVICE FOR THE CONCATENATION OF AUDIOSEGMENTS TAKING INTO ACCOUNT COARTICULATION	

PRELIMINARY AMENDMENT

Assistant Commissioner for Patents and Trademarks
Washington, D.C. 20231
BOX: PCT

Dear Sir:

Prior to calculating the National Stage filing fee, cancel claims 1-68 and insert the following new claims:

1.(new) A method for the co-articulation-specific concatenation of audio segments, in order to generate synthesised acoustical data which reproduces a sequence of concatenated sounds/ phones, comprising the following steps:
- selecting at least two audio segments which contain bands, each of which reproducing a portion of a sound/phone or a portion of a sound/phone sequence,
- establishing a band to be used of an earlier audio segment;
- establishing a band to be used of a later audio segment, which begins with the later audio segment and ends with the co-articulation band of the later audio segment which follows the initially used solo articulation band;

- with the duration and position of the bands to be used being determined as a function of the earlier and later audio segments; and
- concatenating the established band of the earlier audio segment with the established band of the later audio segment, in that the instance of concatenation, as a function of properties of the used band of the later audio segment, is set in a band which begins immediately before the used band of the later audio segment and ends with same.

2. (new) The method according to Claim 1, characterised in that
- the instance of concatenation is set in a band which lies in the vicinity of the boundaries of the initially to be used solo articulation band of the later audio segment, if the band of same to be used reproduces a static sound/phone at the beginning; and
- a downstream portion of the band to be used of the earlier audio segment and an upstream portion of the band to be used of the later audio segment are processed by means of suitable transfer functions and added in an overlapping manner (cross fade), with the transfer functions and the length of an overlapping portion of the two bands being determined depending on the audio segments to be concatenated.

3.(new) The method according to Claim 1, characterised in that
- the instance of concatenation is set in a band which lies immediately before the band to be used of the later audio segment, if the used band of same reproduces a dynamic sound/ phone at the beginning; and
- a downstream portion of the band to be used of the earlier audio segment and an upstream portion of the band to be used of the later audio segment are processed by means of suitable transfer functions and joined in a non-overlapping manner (hard fade), with the transfer functions being determined depending on the acoustical data to be synthesised.

4.(new) The method according to Claim 1 characterised in that for a sound/phone or a portion of the sequence of concatenated sounds/phones at the start of the concatenated sound/phone sequence a band of an audio segment is selected so that the start of the band reproduces the properties of the start of the concatenated sound/phone sequence.

5.(new) The method according to Claim 1 characterised in that for a sound/phone or a portion of the sequence of concatenated sounds/phones at the end of the concatenated sound/phone sequence a band of an audio segment is selected so that the end of the band reproduces the properties of the end of the concatenated sound/phone sequence.

6.(new) The method according to Claim 1 characterised in that the voice data to the synthesised is combined in groups, each of which being described by an individual audio segment.

7.(new) The method according to Claim 1 characterised in that an audio segment is selected for the later audio segment band, which reproduces the highest number of successive portions of the sounds/phones of the sound/phone sequence, in order to use the smallest number of audio segment bands in the generation of the synthesised acoustical data.

8.(new) The method according to Claim 1 characterised in that a processing of the used bands of individual audio segments is carried out by means of suitable functions depending on properties of the concatenated sound/phone sequence, with these properties involving i.a. a modification of the frequency, the duration, the amplitude, or the spectrum.

9.(new) The method according to Claim 1 characterised in that a processing of the used bands of individual audio segments is carried out by means of suitable functions in a band, in which the instance of concatenation lies, with these functions involving i.a. a modification of the frequency, the duration, the amplitude, or the spectrum.

10.(new) The method according to Claim 1 characterised in that the instance of concatenation is set in places of the bands to be used of the earlier and/or later audio segment, in which the two used bands are in agreement with respect to one or several suitable properties, with these properties including i.a.: zero point, amplitude values, gradients, derivatives of any degree, spectra, tone levels, amplitude values within a

frequency band, volume, style of speech, emotion of speech, or other properties covered in the phone classification scheme.

11.(new) The method according to Claim 1 characterised in that
- the selection of the used bands of individual audio segments, their processing, their variation, as well as their concatenation are additionally carried out with the application of heuristic knowledge which is obtained by an additionally carried out heuristic method.

12.(new) The method according to Claim 1 characterised in that
- the acoustical data to be synthesised is voice data, and the sounds are phones.

13.(new) The method according to Claim 2 characterised in that
- the static phones include vowels, diphtongs, liquids, vibrants, fricativen and nasals.

14.(new) The method according to Claim 3 characterised in that and
- the dynamic phones include plosives, affricates, glottal stops, and click sounds.

15.(new) The method according to Claim 1 characterised in that
- a conversion of the synthesised acoustical data to acoustical signals and/or voice signals is carried out.

16.(new) A device for the co-articulation-specific concatenation of audio segments, in order to generate synthesised acoustical data which reproduces a sequence of phones, comprising:
- a database (107) in which audio segments are stored, each of which reproducing portion of a phone or portions of a sequence of (concatenated) phones;
- and/or any upstream synthesis means (108) which supplies audio segments;
- a means (105) for the selection of at least two audio segments from the database (107) and/or the upstream synthesis means (108); and

- a means (111) for the concatenation of audio segments, characterised in that the concatenation means (111) is suited for
- defining a band to be used of an earlier audio segment;
- defining a portion to be used of a later audio segment in a band which starts with the later audio segment and ends after a co-articulation band of the later audio segment, which follows after the initially used solo articulation band;
- determining the duration and position of the used bands depending on the earlier and later audio segments; and
- concatenating the used band of the earlier audio segment with the used band of the later audio segment by defining the instance of concatenation as a function of properties of the used band of the later audio segment in a band which starts immediately before the used band of the later audio segment and ends with same.

17.(new) The device according to Claim 16, characterised in that the concatenation means (111) comprises:
- means for the concatenation of the used band of the earlier audio segment with the used band of the later audio segment, whose used band reproduces a static phone at the beginning in the vicinity of the boundaries of the initially occurring solo articulation band of the used band of the later audio segment;
- means for processing a downstream portion of the used band of the earlier audio segment and an upstream portion of the used band of the later audio segment by suitable transfer functions; and
- means for the overlapping addition of the two bands in an overlapping portion (cross fade), which depends on the audio segments to be concatenated, with the transfer functions and the length of an overlapping portion of the two bands being determined depending on the acoustical data to be synthesised.

18.(new) The device according to Claim 16 characterised in that the concatenation (111) means comprises:

- means for the concatenation of the used band of the earlier audio segment with the used band of the later audio segment, whose used band reproduces a dynamic phone at the beginning, immediately before the used band of the later audio segment;
- means for processing a downstream portion of the used band of the earlier audio segment and an upstream portion of the used band of the later audio segment by suitable transfer functions, with the transfer functions being determined depending on the acoustical data to be synthesised; and
- means for the non-overlapping joining of the two audio segments.

19.(new) The device according to Claim 16 characterised in that the database (107) includes audio segments or the upstream synthesis means (108) supplies audio segments which comprise bands which at the start reproduce a phone or a portion of the concatenated phone sequence at the start of the concatenated phone sequence.

20.(new) The device according to Claim 16 characterised in that the database (107) includes audio segments or the upstream synthesis means (108) supplies audio segments which comprise bands, whose ends reproduce a phone or a portion of the concatenated phone sequence at the end of the concatenated phone sequence.

21.(new) The device according to Claim 16 characterised in that the database (107) includes a group of audio segments or the upstream synthesis means (108) supplies audio segments which comprise bands, whose starts each reproduce only a static phone.

22.(new) The device according to Claim 16 characterised in that the concatenation means (111) comprises:
- means for the generation of further audio segments by concatenation of audio segments, with the starts of the bands each reproducing a static phone, each with a band of a later audio segment whose used band reproduces a dynamic phone at the start, and
- a means which supplies the further audio segments to the database (107) or the selection means (105).

23.(new) The device according to Claim 16 characterised in that, in the selection of the audio segment bands from the database (107) or the upstream synthesis means (108), the selection means (105) is suited to select the audio segments which reproduce the greatest number of successive portions of concatenated phones of the concatenated phone sequence.

24.(new) The device according to Claim 16 characterised in that the concatenation means (111) comprises means for processing the used bands of individual audio segments with the aid of suitable functions, depending on properties of the concatenated phone sequence, with the functions involving among others a modification of the frequency, the duration, the amplitude, or the spectrum.

25.(new) The device according to Claim 16 characterised in that
- the concatenation means (111) comprises means for processing the used bands of individual audio segments with the aid of suitable functions in a band including the instance of concatenation, with this function involving i.a. a modification of the frequency, the duration, the amplitude, or the spectrum.

26.(new) The device according to Claim 16 characterised in that
- the concatenation means (111) comprises means for the selection of the instance of concatenation in a place in the used bands of the earlier and/or the later audio segment, in which the two used bands are in agreement with respect to one or several suitable properties, with these properties including i.a.: zero points, amplitude values, gradients, derivatives of any degree, spectra, tone levels, amplitude values in a frequency band, volume, style of speech, emotion of speech, or other properties covered in the phone classification scheme.

27.(new) The device according to Claim 16 characterised in that
- the selection means (105) comprises means for the implementation of heuristic knowledge which relates to the selection of the used bands of the individual audio segments, their processing, their variation, as well as their concatenation.

28.(new) The device according to Claim 16 characterised in that
- the database (107) includes audio segments or the upstream synthesis means (108) supplies audio segments which include bands, each of which reproducing at least a portion of a sound or phone, respectively, a sound or phone, respectively, portions of phone sequences or polyphones, respectively, or sound sequences or polyphones, respectively.

29.(new) The device according to Claim 17 characterised in that
the data base (107) includes audio segments or the upstream synthesis means (108) supplies audio segments, with a static sound corresponding to a static phone and comprising vowels, diphtongs, liquids, vibrants, fricatives, and nasals.

30.(new) The device according to Claim 18 characterised in that
- the database (107) includes audio segments or the upstream synthesis means (108) supplies audio segments, with a dynamic sound corresponding to a dynamic phone and comprising plosives, affricates, glottal stops, and klick speech.

31.(new) The device according to Claim 16 characterised in that
- the concatenation means (111) is suitable to generate synthesised voice data by means of the concatenation of audio segments.

32.(new) The device according to Claim 16 characterised in that
- means (117) are provided for the conversion of the synthesised acoustical data to acoustical signals and/or voice signals.

33.(new) A data carrier which includes a computer program for the co-articulation-specific concatenation of audio segments in order to generate synthesised acoustical data which reproduces a sequence of concatenated phones, comprising the following steps:
- selection of at least two audio segments which contain bands, each of which reproducing a portion of a sound/phone or a portion of a sound/phone sequence,

characterised by the steps of:

- establishing a band to be used of an earlier audio segment;
- establishing a band to be used of a later audio segment, which begins with the later audio segment and ends with the co-articulation band of the later audio segment which follows the initially used solo articulation band;
- with the duration and position of the bands to be used being determined as a function of the earlier and later audio segments; and
- concatenating the established band of the earlier audio segment with the established band of the later audio segment, in that the instance of concatenation, as a function of properties of the used band of the later audio segment, is set in its established band which starts immediately before the band to be used of the later audio segment and ends with same.

34.(new) The data carrier according to Claim 33, characterised in that the computer program selects the instance of the concatenation of the used band of the second audio segment with the used band of the first audio segment in such a manner that

- the instance of concatenation is set in a band which lies in the vicinity of the boundaries of the initially used solo articulation band of the later audio segment, if its used band reproduces a static phone at the start;
- a downstream portion of the used band of the earlier audio segment and an upstream portion of the used band of the later audio segment are processed by suitable transfer functions and added in an overlapping manner (cross fade), with the transfer functions and the length of an overlapping portion of the two bands being determined depending on the audio segments to be concatenated.

35.(new) The data carrier according to Claim 33 characterised in that the computer program selects the instance of the concatenation of the used band of the second audio segment with the used band of the first audio segment in such a manner that

- the instance of concatenation is set in a band which lies immediately before the used band of the later audio segment, if its used band reproduces a dynamic phone at the start;

- a downstream portion of the used band of the earlier audio segment and an upstream portion of the used band of the later audio segment are processed by suitable transfer functions and added in a non-overlapping manner (hard fade), with the transfer functions being determined depending on the audio segments to be concatenated.

36.(new) The data carrier according to Claim 33 characterised in that the computer program selects a band of an audio segment for a phone or a portion of the sequence of concatenated phones at the start of the concatenated phone sequence, the start of which reproduces the properties of the start of the concatenated sequence of phones.

37.(new) The data carrier according to Claim 33 characterised in that the computer program selects a band of an audio segment for a phone or a portion of the sequence of concatenated phones at the end of the concatenated phone sequence, the end of which reproduces the properties of the end of the concatenated sequence of phones.

38.(new) The data carrier according to Claim 33 characterised in that the computer program carries out a processing of the used bands of individual audio segments with the aid of suitable functions depending on properties of the phone sequence, with the functions involving i.a. modification of the frequency, the duration, the amplitude, or the spectrum.

39.(new) The data carrier according to Claim 33 characterised in that the computer program selects an audio segment band for the later audio segment band which reproduces the highest number of successive portions of the concatenated phones in the phone sequence, in order to use the smallest number of audio segment bands in the generation of the synthesised acoustical data.

40.(new) The data carrier according to Claim 39 characterised in that the computer program carries out a processing of the used bands of individual audio segments with the aid of suitable functions in a band in which the instance of concatenation lies, with these

functions involving i.a. a modification of the frequency, the duration, the amplitude, or the spectrum.

41.(new) The data carrier according to Claim 33 characterised in that the computer program establishes the instance of concatenation in a place of the used bands of the first and/or the second audio segment, in which the two used bands are in agreement with respect to one or several suitable properties, with these properties including i.a.: zero points, amplitude values, gradients, derivatives of any degree, spectra, tone levels, amplitude values in a frequency band, volume, style of speech, emotion of speech, or other properties covered in the phone classification scheme.

42.(new) The data carrier according to Claim 33 characterised in that the computer program carries out an implementation of heuristic knowledge which relates to the selection of the used bands of the individual audio segments, their processing, their variation, as well as their concatenation.

43.(new) The data carrier according to Claim 33 characterised in that the computer program is suited for the generation of synthesised voice data, with the sounds being phones.

44.(new) The data carrier according to Claim 34 characterised in that the computer program is suited for the generation of static phones, with the static phones comprising vowels, diphtongs, liquids, vibrants, fricatives, and nasals.

45.(new) The data carrier according to Claim 35 characterised in that the computer program is suited for the generation of dynamic phones, with the dynamic phones comprising plosives, affricates, glottal stops, and klick speech.

46.(new) The data carrier according to Claim 33 characterised in that the computer program converts the synthesised acoustical data to acoustical convertible data and/ or voice signals.

47.(new) Synthesised voice signals which consist of a sequence of sounds or phones, respectively, with the voice signals being generated in that:

- at least two audio segments are selected which reproduce the sounds or phones, respectively; and
- the audio segments are linked by a co-articulation-specific concatenation, with
- one band to be used of an earlier audio segment being established;
- one band to be used of a later audio segment being established which starts with the later audio segment and ends with the co-articulation band of the later audio segment, following the initially used solo articulation band;
- with the duration and position of the bands to be used being determined depending on the audio segments; and
- the used bands of the audio segments being concatenated in a co-articulation-specific manner, in that the instance of concatenation, as a function of properties of the used band of the later audio segment, is set in a band which starts immediately before the used band of the later audio segment and ends with same.

48.(new) The synthesised voice signals according to Claim 47, characterised in that the voice signals are generated in that

- the audio segments are concatenated in an instance which lies in the vicinity of the boundaries of the later audio segment, if the start of this band reproduces a static sound or phone, respectively, with the static phone being a vowel, a diphtong, a liquid, a fricative, a vibrant, or a nasal; and
- a downstream portion of the used band of the earlier audio segment and an upstream portion of the used band of the later audio segment are processed by means of suitable transfer function and both bands are added in an overlapping manner (cross fade), with the transfer functions and the length of an overlapping portion of the two bands being determined depending on the audio segments to be concatenated.

49.(new) The synthesised voice signals according to Claim 47 characterised in that the voice signals are generated in that

- the audio segments are concatenated in an instance which lies immediately before the used band of the later audio segment, if the start of this band reproduces a dynamic sound or phone, respectively, with the dynamic phone being a plosive, an affricate, a glottal stop, or klick speech; and
- a downstream portion of the used band of the earlier audio segment and an upstream portion of the used band of the later audio segment are processed by means of suitable transfer functions and both bands are joined in a non-overlapping manner (hard fade), with the transfer functions being determined depending on the audio segments to be concatenated.

50.(new) The synthesised voice signals according to Claim 47 characterised in that
- the first sound or the first phone, respectively, or a portion of the first phone sequence or of the first polyphone, respectively, in the sequence is generated by an audio segment, whose used band at the start reproduces the properties of the start of the sequence.

51.(new) The synthesised voice signals according to Claim 47 characterised in that
- the last sound or the last phone, respectively, or a portion of the last phone sequence or of the last polyphone, respectively, in the sequence is generated by an audio segment, whose used band at the end reproduces the properties of the end of the sequence.

52.(new) The synthesised voice signals according to Claim 47 characterised in that
- the voice signals are generated in that later bands of audio segments, beginning with the reproduction of a dynamic sound or phone, respectively, are concatenated with earlier bands of audio segments, beginning with the reproduction of a static sound or phone, respectively.

53.(new) The synthesised voice signals according to Claim 47 characterised in that
- such audio segments are selected which reproduce the highest number of portions of sounds or phones, respectively, of the sequence, in order to use the smallest number of audio segment bands in the generation of the voice signals.

54.(new) The synthesised voice signals according to Claim 47 characterised in that - the voice signals are generated by the concatenation of the used bands of audio segments which are processed with the aid of suitable functions depending on properties of the sound sequence or phone sequence, respectively, with the functions involving i.a. a modification of the frequency, the duration, the amplitude, or the spectrum.

55.(new) The synthesised voice signals according to Claim 47 characterised in that - the voice signals are generated by the concatenation of the used bands of audio segments which are processed with the aid of suitable functions depending on properties of the sound sequence or phone sequence, respectively, in an area in which the instance of concatenation lies, with these properties including i.a. a modification of the frequency, the duration, the amplitude, or the spectrum.

56.(new) The synthesised voice signals according to Claims 47 characterised in that the instance of concatenation lies at a place in the used bands of the earlier and/or the later audio segment, in which the two used bands are in agreement with respect to one or several suitable properties, with these properties including i.a.: zero points, amplitude values, gradients, derivatives of any degree, spectra, tone levels, amplitude values in a frequency band, volume, style of speech, emotion of speech, or other properties covered in the phone classification scheme.

57.(new) The synthesised voice signals according to Claim 47 characterised in that the voice signals are suited for a conversion to acoustic signals.

58.(new) An acoustical, optical, magnetic, or electrical data storage which contains audio segments in order generate synthesised acoustical data by means of a concatenation of used bands of the audio segments, utilising the methods according to Claim 1.

59.(new) The data storage according to Claim 58, characterised in that a group of the audio segments reproduces sounds or phones, respectively, or portions of sounds or phones, respectively.

60.(new) The data storage according to Claim 58 characterised in that a group of the audio segments reproduces phone sequences or portions of phone sequences or polyphones, respectively, or portions of polyphones.

61.(new) The data storage according to Claim 58 characterised in that a group of audio segments is provided whose used bands start with a static sound or phone, respectively, with the static phones comprising vowels, diphtongs, liquids, fricatives, vibrants, and nasals.

62.(new) The data storage according to Claim 58 characterised in that audio segments are provided which are suitable for the conversion to acoustical signals

63.(new) The data storage according to Claim 58 which additionally contains information in order to carry out a processing of the used bands of individual audio segments with the aid of suitable functions depending on properties of the acoustical data to be synthesised, with the functions involving i.a. a modification of the frequency, the duration, the amplitude, or the spectrum.

64.(new) The data storage according to Claim 58 which additionally contains information relating to a processing of the used bands of individual audio segments with the aid of suitable functions in a band in which the instance of concatenation lies, with this function involving i.a. a modification of the frequency, the duration, the amplitude, or the spectrum.

65.(new) The data storage according to Claim 58 which additionally provides linked audio segments, whose instance of concatenation lies at a place of the used bands of the earlier and/or later audio segment, where both used bands are in agreement with respect to one or several suitable properties with these properties being i.a.: zero points, amplitude values, gradients, derivatives of any degree, spectra, tone levels, amplitude values

in a frequency band, volume, style of speech, emotion of speech, or other properties covered in the phone classification scheme.

66.(new) The data storage according to Claim 51,
which additionally contains information in the form of heuristic knowledge, which relates to the selection of the used bands of the individual audio segments, their processing, their variation, as well as their concatenation.

Respectfully submitted,

2/17/01
Date

Thomas R. FitzGerald
Registration No. 26,730

JAECKLE FLEISCHMANN & MUGEL, LLP
39 State Street
Rochester, New York 14614-1310
Telephone: (716) 262-3640
Facsimile: (716) 262-4133

116355

VERIFICATION OF A TRANSLATION

I, the below named translator, hereby declare that:

My name and post office address are as stated below:

Brita Baumgärtel
Mittermayrstr. 12
D-80796 München

I am knowledgeable about the English language and about the language in which the below identified international application was filed, and I believe the English translation of the international application No. PCT/EP 99/06081 is a true and complete translation of the above international application as filed.

I hereby declare that all statements made therein of my own knowledge are true and that all statements made on information and belief are believed to be true; and further that these statements were made with the knowledge that willful false statements and the like so made are punishable by fine or imprisonment, or both, under Section 1001 of Title 18 of the United States Code and that such willful false statements may jeopardize the validity of the application or any patent issued thereon.

Munich, February 5, 2001

Brita Baumgärtel
.....
(Translator)

Method and Devices for the Co-articulation-specific Concatenation of Audio Segments

The invention relates to a method and a device for the concatenation of audio segments for the generation of synthesised acoustical data, in particular synthesised speech. In particular, the invention relates to synthesised voice signals which have been generated by the inventive co-articulation-specific concatenation of voice segments, as well as to a data carrier which contains a computer program for the inventive generation of synthesised acoustical data, in particular, synthesised speech.

In addition, the invention relates to a data storage which contains audio segments which are suited for the inventive co-articulation-specific concatenation, and a sound carrier which, according to the invention, contains synthesised acoustical data.

It must be emphasised that both the state of the art represented in the following, and the present invention relate to the entire field of the synthesis of acoustical data by means of the concatenation of individual audio segments which are obtained in any manner. However, for the sake of simplifying the discussion of the state of the art as well as the description of the present invention, the following explanations refer specifically to synthesised voice data by means of the concatenation of individual voice segments.

During the past years, the data-based approach has been successful over the rule-based approach in the field of speech synthesis, and can be found in various methods and systems for speech synthesis. Although the rule-based approach principally enables a better speech synthesis, it is necessary for its implementation to explicitly phrase the entire knowledge which

is required for speech generation, i.e. to formally model the speech to be synthesised. Due to the fact that the known speech models comprise a simplification of the speech to be synthesised, the voice quality of the speech generated in this manner is not sufficient.

For this reason, a data-based speech synthesis is carried out to an increasing extent, wherein corresponding segments are selected from a database containing individual voice segments and linked (concatenated) to each other. In this context, the voice quality is primarily depending on the number and type of the available voice segments, because only that speech can be synthesised which is reproduced by voice segments in the database. In order to minimise the number of the voice segments to be provided and, nevertheless, to still generate a high quality synthesised speech, various methods are known which carry out a linking (concatenation) of the voice segments according to complex rules.

When using such methods or corresponding devices, respectively, an inventory, i.e. a database comprising the voice audio segments can be employed which is complete and manageable. An inventory is complete if it is capable of generating any sound sequence of the speech to be synthesised, and it is manageable if the number and type of the data of the inventory can be processed in a desired manner by means of the technically available means. Furthermore, such a method must ensure that the concatenation of the individual inventory elements generates a synthesised speech which differs as little as possible from a naturally spoken speech. To this end, a synthesised speech must be fluent and comprise the same articulatory effects as a natural speech. In this context, the so-called co-articulatory effects, i.e. the mutual influence of phones, are of particular importance. For this reason, the inventory elements should be of such a nature that they consider the co-

articulation of individual successive phones. In addition, a method for the concatenation of the inventory elements should link the elements, even beyond word and phrase boundaries, under consideration of the co-articulation of individual successive phones as well as of the higher-order co-articulation of several successive phones.

Before presenting the state of the art, a few terms from the field of speech synthesis, which are necessary for a better understanding, will be explained in the following:

- A phone is a class of any sound events (noises, sounds, tones, etc.). The sound events are classified in accordance with a classification scheme into phone classes. A sound event belongs to a phoneme if the values of the sound event are within the range of values defined for the phone with respect to the parameters (e.g. spectrum, tone level, volume, chest or head voice, co-articulation, resonance cavities, emotion, etc.) used for the classification.
The classification scheme for phones depends on the type of application. For vocal sounds (= phones), the IPA classification is generally used. However, the definition of the term phone as used herein is not limited to this, but any other parameters can be used. If, for example, in addition to the IPA classification, the tone level or the emotional expression are included as parameters in the classification, two 'a' phones with different tone level or different emotional expression become different phones in the sense of the definition. Phones can, however, also be the tones of a musical instrument, e.g. a violin, in the different tone levels and the different modes of playing (up-bow and down-bow, detaché, spiccato, marcato, pizzicato, col legno, etc.). Phones can be the barking of dogs or the squealing of a car door.

Phones can be reproduced by audio segments which contain corresponding acoustical data.

In the description of the invention following the definitions, the term vocal sound can invariably be replaced by the term phone in the sense of the previous definition, and the term phoneme can be replaced by the term phonetic character. (This also applies the other way round, because phones are vocal sounds classified according to the IPA classification).

- A static phone has bands which are similar to previous or subsequent bands of the static phone. The similarity need not necessarily be an exact correspondence as in the periods of a sinusoidal tone, but is analogous to the similarity as it prevails between the bands of the static phones defined in the following.

- A dynamic phone has no bands with a similarity with previous or subsequent bands of the dynamic phone, such as, e.g. the sound event of an explosion or a dynamic phone.

- A phone is a vocal sound which is generated by the organs of speech (a vocal sound). The phones are classified into static and dynamic phones.

- The static phones include vowels, diphtongs, nasals, laterals, vibrants, and fricatives.

- The dynamic phones include plosives, affricates, glottal stops, and click sounds.

- A phoneme is the formal description of a phone, with the formal description usually being effected by phonetic characters.

- The co-articulation refers to the phenomenon that a sound, i.e. a phone, too, is influenced by upstream or downstream sounds or phones, respectively, with the co-articulation occurring both between immediately neighbouring sounds/phones, but also covering a sequence of several sounds/phones as well (for example in rounding the lips).

A sound or phone, respectively, can therefore be classified into three bands (see also Fig. 1b):

- The initial co-articulation band comprises the band from the start of a sound/phone to the end of the co-articulation due to a upstream sound/phone.

- The solo articulation band is the band of the sound/phone which is not influenced by an upstream or downstream sound or an upstream or downstream phone, respectively.

- The end co-articulation band comprises the band from the start of the co-articulation due to a downstream sound/phone to the end of the sound/phone.

- The co-articulation band comprises an end co-articulation band and the neighbouring initial co-articulation band of the neighbouring sound/phone.

- A polyphone is a sequence of phones.

- The elements of an inventory are audio segments stored in a coded form which reproduce sounds, portions of sounds, sequences of sounds, or portions of sequences of sounds, or phones, portions of phones, polyphones, or portions of polyphones, respectively. For a better understanding of the potential structure of an audio segment/inventory element, reference is made to Fig. 2a which shows a conventional audio

segment, and Figs. 2b - 21 which show inventive audio segments. In addition, it should be mentioned that audio segments can can be formed from smaller or larger audio segments which are included in the inventory or a database. Furthermore, audio segments can also be provided in a transformed form (e.g. in a Fourier-transformed form) in the inventory or the database. Audio segments for the present invention can also come from a prior synthesis step (which is not part of the method). Audio segments include at least a part of an initial co-articulation band, a solo articulation band, and/or an end co-articulation band. In lieu of audio segments, it is also possible to use bands of audio segments.

- The term concatenation implies the joining of two audio segments.

- The concatenation instance if the point of time in which two audio segments are joined.

The concatenation can be effected in various ways, e.g. with a cross fade or a hard fade (see also Figs. 3a - 3e):

- In a cross fade, a downstream band of a first audio segment band and an upstream band of a second audio segment band are processed by means of suitable transfer functions, and subsequently these two bands are overlappingly added in such a manner that at the most the shorter band with respect to time of the two bands is completely overlapped by the longer one with respect to time of the two band.

- In a hard fade, a later band of a first audio segment and an earlier band of a second audio segment are processed by means of suitable transfer functions, with the two audio segments being joined to one another in such a manner that the later

band of the first audio segment and the earlier band of the second audio segment do not overlap.

The co-articulation band is primarily noticeable in that a concatenation therein is associated with discontinuities (e.g. spectral skips).

In addition, reference is to be made that, strictly speaking, a hard fade is a boundary case of a cross fade, in which an overlap of a later band of a first audio segment and an earlier band of a second audio segment has a length of zero. This allows to replace a cross fade with a hard fade in certain, e.g. extremely time-critical applications, with such an approach to be contemplated scrupulously, because it results in considerable quality losses in the concatenation of audio segments which actually are to be concatenated by a cross fade.

- The term prosody refers to changes in the voice frequency and the voice rhythm which occur in spoken words or phrases, respectively. The consideration of such prosodic information is necessary in the speech synthesis in order to generate a natural word or phrase melody, respectively.

From WO 95/30193 a method and a device are known for the conversion of text to audible voice signals under utilising a neural network. For this purpose, the text to be converted to speech is converted to a sequence of phonema by means of a converter unit, with information on the syntactic boundaries of the text and the stress of the individual components of the text being additionally generated. This information, together with the phonema, are transferred to a device which determines the duration of the pronunciation of the individual phonema in a rule-based manner. A processor generates a suitable input for the neural network from each individual phoneme in connec-

tion with the corresponding syntactic and time-related information, with said input for the neural network also comprising the corresponding prosodic information for the entire phoneme sequence. From the available audio segments the neural network then selects only those segments which best reproduce the input phonema and links said audio segments accordingly. In this linking operation the individual audio segments with respect to their duration, total amplitude, and frequency are matched to upstream and downstream audio segments under consideration of the prosodic information of the speech to be synthesised and time successively connected with each other. A modification of individual bands of the audio segments is not described therein.

For the generation of the audio segments which are required for this method, the neural network has first to be trained by dividing naturally spoken speech into phones or phone sequences and assigning these phones or phone sequences corresponding phonema or phoneme sequences in the form of audio segments. Due to the fact that this method provides for a modification of individual audio segments only, but not for a modification of individual bands of an audio segment, the neural network must be trained with as many different phones or phone sequences as possible for converting any text to a synthesised speech with a natural sound. Depending of the application, this may prove to require very high expenditures. On the other hand, an insufficient training process of the neural network may have a negative influence on the quality of the speech to be synthesised. Moreover, it is not possible with the method described therein to determine the concatenation instance of the individual audio segments depending on upstream or downstream audio segments, in order to perform a co-articulation-specific concatenation.

US-5,524,172 describes a device for the generation of synthesised speech, which utilises the so-called diphone method. Here, a text which is to be converted to synthesised speech is divided into phoneme sequences, with corresponding prosodic information being assigned to each phoneme sequence. From a database which contains audio segments in the form of diphones, for each phoneme of the sequence two diphones reproducing the phoneme are selected and concatenated under consideration of the corresponding prosodic information. In the concatenation the two diphones each are weighted by means of a suitable filter, and the duration and tone level of both diphones modified in such a manner that upon the linking of the diphones a synthesised phone sequence is generated, whose duration and tone level correspond to the duration and tone level of the desired phoneme sequence. In the concatenation the individual diphones are added in such a manner that a later band of a first diphone and an earlier band of a second diphone overlap, with the instance of concatenation being generally in the area of stationary bands of the individual diphones (see Fig. 2a). Due to the fact that a variation of the instance of concatenation under consideration of the coarticulation of successive audio segments (diphones) is not intended, the quality (naturalness and audibility) of a speech synthesised in such a manner can be negatively influenced.

A further development of the previously discussed method can be found in EP-0,813,184 A1. In this case, too, a text to be converted to synthesised speech is divided into individual phonema or phoneme sequences, and corresponding audio segments are selected from a database and concatenated. In order to achieve an improvement of the synthesised speech, two approaches have been realised with this method, which differ from the state of the art discussed so far. With the use of a smoothing filter which accounts for the lower-frequency harmonic frequency components of an upstream and a downstream

audio segment, the transition from the upstream audio segment to the downstream audio segment is to be optimised, in that a later band of the upstream audio segment and an earlier band of the downstream audio segment in the frequency range are tuned to each other. In addition, the database provides audio segments which are slightly different from one another but are suited for synthesising one and the same phoneme. In this manner, the natural variation of the speech is to be mimicked in order to achieve a higher quality of the synthesised speech. Both the use of the smoothing filter and the selection from a plurality of various audio segments for the realisation of a phoneme require a high computing power of the used system components in the implementation of this method. Moreover, the volume of the database increases due to the increased number of the provided audio segments. Furthermore, this method, too, does not provide for a ca-articulation dependent choice of the concatenation instance of individual audio segments, which may reduce the quality of the synthesised speech.

DE 693 18 209 T2 deals with formant synthesis. According to this document two multi-voice phones are connected with each other using an interpolation mechanism which is applied to a last phoneme of an upstream phone and to a first phoneme of a downstream phone, with the two phonema of the two phones being identical and with the connected phones are superposed to one phoneme. Upon the superposition, each of the curves describing the two phonema is weighted with a weighting function. The weighting function is applied to a band of each phoneme, which begins immediately after the start of the phoneme and ends immediately before the end of the phoneme. Thus, in the concatenation of phones described therein, the bands of the phonema, which form the transition between phones, correspond essentially to the respective entire phonema. This means, that portions of the phonema used for concatenation, invariably comprise all three bands, i.e. the respective initial co-

articulation band, solo articulation band, and end co-articulation band. Consequently, D1 teaches an approach how the transitions between two phones are to be smoothed.

Moreover, according to this document the instance of the concatenation of two phones is established in such a manner that the last phoneme in the upstream phone and the first phoneme in the downstream phone completely overlap.

Principally, it is to be stated that DE 689 15 353 T2 aims at improving the tone quality, in that an approach is specified how to design the transition between two neighbouring sampling values. This is of particular relevance in the case of low sampling rates.

In the speech synthesis described in this document, waveforms are used which reproduce the phones to be concatenated. With waveforms for upstream phones, a corresponding final sampling value and an associated zero crossing point are established, while with waveforms for downstream phones, a corresponding first upper sampling value and an associated zero crossing point are established. Depending on these established sampling values and the associated zero crossing points, phones are connected with each other by means of maximal four different ways. The number of connection types is reduced to two, if the waveforms are generated by utilising the Nyquist theoreme. DE 689 15 353 T2 describes that the used band of waveforms extends between the last sampling value of the upstream waveform and the first sampling value of the downstream waveform. A variation of the duration of the used bands as a function of the waveforms to be concatenated, as it is the case with the invention, is not disclosed in D1.

In summary, it can be said that the state of the art allows to synthesise any phoneme sequences, but that the phoneme se-

quences synthesised in this manner do not possess an authentic voice quality. A synthesised phoneme sequence has an authentic voice quality if it cannot be distinguished by a listener from the same phoneme sequence spoken by a real speaker.

Methods are also known which use an inventory which comprises complete words and/or phrases in authentic voice quality as inventory elements. For the speech synthesis, these elements are brought into a desired order, with the possibilities of various voice sequences being limited to a high degree by the volume of such an inventory. The synthesis of any phoneme sequences is not possible with these methods.

It is therefore the object of the present invention to provide a method and a corresponding device which eliminate the problems of the state of the art and enable the generation of synthesised acoustical data, in particular, synthesised voice data, which a listener cannot distinguish from corresponding natural acoustical data, in particular, naturally spoken speech. The acoustical data synthesised by means of the invention, in particular, synthesised voice data, is to possess an authentic acoustical quality, in particular, an authentic voice quality.

For the solution of this object the invention provides a method according to Claim 1, a device according to Claim 14, synthesised voice signals according to Claim 28, a data carrier according to Claim 39, a data storage according to Claim 51, as well as a sound carrier according to Claim 60. The invention therefore makes it possible to generate synthesised acoustical data which reproduces a sequence of phones, in that in the concatenation of audio segments, the instance of the concatenation of two audio segments is determined, depending on properties of the audio segments to be linked, in particular the co-articulation effects which relate

to the two audio segments. According to the present invention, the instance of concatenation is preferably selected in the vicinity of the boundaries of the solo articulation band. In this manner, a voice quality is achieved, which cannot be obtained with the state of the art. The required computation power is not higher than with the state of the art.

In order to mimic the variations which can be found in the corresponding natural acoustical data, in the synthesis of acoustical data, the invention provides for a different selection of the audio segment bands as well as for different ways of the co-articulation-specific concatenation. A higher degree of naturalness of the synthesised acoustical data is achieved if a later audio segment band, whose start reproduces a static phone, is connected with an earlier audio segment band by means of a cross fade, or if a later audio segment band, whose start reproduces a dynamic phone, is connected with an earlier audio segment band by means of a hard fade, respectively. In addition, it is advantageous to generate the start of the synthesised acoustical data to be generated by using an audio segment band which reproduces the start of a phone sequence, or to generate the end of the synthesised acoustical data to be generated by using an audio segment band which reproduces the end of a phone sequence, respectively.

In order to carry out the generation of the synthesised acoustical data in a simpler and faster way, the invention makes it possible to reduce the number of audio segment bands which are required for data synthesising, in that audio segment bands are used which always start with the reproduction of a dynamic phone, which allows to carry out all concatenations of these audio segment bands by means of a hard fade. For this purpose, later audio segment bands are connected with earlier audio segment bands whose starts always reproduce a dynamic phone. In this manner, high-quality synthesised

acoustical data according to the invention can be generated with low computing power (e.g. in the case of answering machines or car navigation systems).

In addition, the invention provides for mimicking acoustical phenomena which result because of a mutual influence of individual segments of corresponding natural acoustical data. In particular, it is intended here to process individual audio segments or individual bands of the audio segments, respectively, with the aid of suitable functions. Thus it is possible to modify i.a. the frequency, the duration, the amplitude, or the spectrum of the audio segments. If synthesised voice data is generated by means of the invention, then preferably prosodic information and/or higher-order co-articulation effects are taken into consideration for the solution of this object.

The signal characteristic of synthesised acoustical data can additionally be improved if the concatenation instance is set in places of the individual audio segment bands to be connected, where the two used bands are in agreement with each other with respect to one or several suitable properties. These properties can be i.a.: zero point, amplitude value, gradient, derivative of any degree, spectrum, tone level, amplitude value in a frequency band, volume, style of speech, emotion of speech, or other properties covered in the phone classification scheme.

The invention further enables to improve the selection of audio segment bands for the generation of the synthesised acoustical data, as well as to make their concatenation more efficient, in that heuristic knowledge is used which relates to the selection, processing, variation, and concatenation of the audio segment bands.

In order to generate synthesised acoustical data which is voice data which does not differ from corresponding natural voice data, preferably audio segment bands are used which reproduce sounds/phones or portions of sound sequences/phone sequences.

Furthermore, the invention permits the utilisation of the generated synthesised acoustical data, in that this data is convertible to acoustical signals and/or voice signals, and/or storable in a data carrier.

In addition, the invention can be used for providing synthesised voice signals which differ from known synthesised voice signals in that, concerning their naturalness and audibility, they do not differ from real speech. For this purpose, audio segment bands are concatenated in a co-articulation-specific manner, each of which reproduces portions of the sound sequence/phone sequence of the speech to be synthesised, in that the bands of the audio segments to be used as well as the instance of the concatenation of these band are established according to the invention as defined in Claim 28.

A further improvement of the synthesised speech can be achieved if a later audio segment band whose start reproduces a static phone is connected with an earlier audio segment band by means of a cross fade, or if a later audio segment band whose start reproduces a dynamic phone, respectively, is connected with an earlier audio segment band by means of a hard fade. Herein, static phones comprise vowels, diphtongs, liquids, fricatives, vibrants, and nasals, and dynamic phones comprise plosives, affricates, glottal stops, and klick speech.

Due to the fact that the start and end stresses of phones in a natural speech differ from comparable, but embedded phones, it

is to be preferred to use corresponding audio segment bands, whose starts reproduce the start of the speech to be synthesised and whose ends reproduce the end of same, respectively.

In particular in the generation of synthesised speech, a fast and efficient procedure is desirable. For this purpose, it is to be preferred to carry out the inventive co-articulation-specific concatenation invariably by means of hard fades, with only such audio segment bands being used whose starts always reproduce a dynamic sound or phone, respectively. Such audio segment bands can be generated in advance according to the invention by means of the co-articulation-specific concatenation of corresponding audio segment bands.

In addition, the invention provides voice signals which have a natural flow of speech, speech melody, and speech rhythm, in that audio segment bands are processed before and/or after the concatenation in their entirety or in individual bands by means of suitable functions. It is particularly advantageous to perform this variation additionally in areas in which the corresponding instances of concatenation are set in order to change i.a. the frequency, duration, amplitude, or spectrum.

An still further improved signal characteristic can be achieved if the concatenation instances are set in places of the audio segment bands to be linked, where these are in agreement with respect to one or several properties.

In order to permit a simple utilisation and/or further processing of the inventive voice signals by means of known methods or devices, such as a CD player, it is to be preferred in particular that the voice signals are convertible to acoustical signals or are storable in a data carrier.

For the purpose of applying the invention also to known devices such as a personal computer or a computer-controlled musical instrument, a data carrier is provided which contains a computer program which enables the performance of the inventive method or the control of the inventive device and its various embodiments, respectively. In addition, the inventive data carrier also permits the generation of voice signals which comprise co-articulation-specific concatenations.

For providing an inventory comprising audio segments, by means of which synthesised acoustical data, in particular synthesised voice data, can be generated which does not differ from corresponding natural acoustical data, the invention provides a data storage which includes audio segments which are suited for being inventively concatenated to synthesised acoustical data. Preferably, such a data carrier includes audio segments which are suited for the performance of the inventive method, for application in the inventive device, or the inventive data carrier. Alternatively, the data carrier can also include inventive voice signals.

In addition, the invention makes it possible to provide inventive synthesised acoustical data, in particular synthesised voice data, which can be utilised with conventional devices, e.g. a tape recorder, a CD player, or a PC audio card. For this purpose, a sound carrier is provided which comprises data which at least partially has been generated by the inventive method or by means of the inventive device or by using the inventive data carrier or the inventive data storage, respectively. The sound carrier may also comprise data which are the inventively co-articulation-specific concatenated voice signals.

Further properties, characteristics, advantages, or modifications of the invention will be explained with reference to the following description; in which:

Fig. 1a is a schematic representation of an inventive device for the generation of synthesised acoustical data;
Fig. 1b shows the structure of a sound/phone;
Fig. 2a shows the structure of a conventional audio segment according to the state of the art, consisting of portions of two phones, i.e. a diphone for voice. It is essential that the solo articulation bands each are included only partially in the conventional diphone audio segment.
Fig. 2b shows the structure of an inventive audio segment which reproduces portions of a sound/phone with downstream co-articulation bands (for voice a quasi 'displaced' diphone);
Fig. 2c shows the structure of an inventive audio segment which reproduces portions of a sound/phone with upstream co-articulation bands;
Fig. 2d shows the structure of an inventive audio segment which reproduces portions of a sound/phone with downstream co-articulation bands and includes additional bands;
Fig. 2e shows the structure of an inventive audio segment which reproduces portions of a sound/phone with upstream co-articulation bands and includes additional bands;
Fig. 2f shows the structure of an inventive audio segment which reproduces portions of several sounds/phones (for speech: a polyphone) with downstream co-articulation bands each. The sounds/phones 2 to (n-1) each are completely included in the audio segment.
Fig. 2g shows the structure of an inventive audio segment which reproduces portions of several sounds/phones (for speech: a polyphone) with upstream co-articulation bands each. The sounds/phones 2 to (n-1) each are completely included in the audio segment.

Fig. 2h shows the structure of an inventive audio segment which reproduces portions of several sounds/phones (for speech: a polyphone) with downstream co-articulation bands each and includes additional bands. The sounds/phones 2 to (n-1) each are completely included in the audio segment.

Fig. 2i shows the structure of an inventive audio segment which reproduces portions of several sounds/phones (for speech: a polyphone) with downstream co-articulation bands each and includes additional bands. The sounds/phones 2 to (n-1) each are completely included in the audio segment.

Fig. 2j shows the structure of an inventive audio segment which reproduces a portion of a sound/phone of the start of a sound sequence/phone sequence;

Fig. 2k shows the structure of an inventive audio segment which reproduces portions of sounds/phones of the start of a sound sequence/phone sequence;

Fig. 2l shows the structure of an inventive audio segment which reproduces a sound/phone of the end of a sound sequence /phone sequence;

Fig. 3a shows the concatenation according to the state of the art by means of an example of two conventional audio segments. The segments begin and end with portions of the solo articulation bands (generally half of same).

Fig. 3aI shows the concatenation according to the state of the art. The solo articulation band of the middle phone comes from two different audio segments.

Fig. 3b shows the concatenation according to the inventive method by means of an example of two audio segments, each of which containing a sound/phone with downstream co-articulation bands. Both sounds/phones come from the centre of a phone unit sequence.

Fig. 3bI shows the concatenation of these audio segments by means of a cross fade.

The solo articulation band comes from an audio segment. The transition between the audio segments is effected between two

bands and is therefore less susceptible to variations (in spectrum, frequency, amplitude, etc.). The audio segments can also be processed by means of additional transfer functions prior to the concatenation.

Fig. 3bII shows the concatenation of these audio segments by means of a hard fade;

Fig. 3c shows the concatenation according to the inventive method by means of an example of two inventive audio segments, each of which containing a sound/phone with downstream co-articulation bands, with the first audio segment coming from the start of a phone sequence.

Fig. 3cI shows the concatenation of these audio segments by means of a cross fade;

Fig. 3cII shows the concatenation of these audio segments by means of a hard fade;

Fig. 3d shows the concatenation according to the inventive method by means of an example of two inventive audio segments, each of which containing a sound/phone with upstream co-articulation bands. Both audio segments come from the centre of a phone sequence.

Fig. 3dI shows the concatenation of these audio segments by means of a cross fade. The solo articulation band comes from an audio segment.

Fig. 3dII shows the concatenation of these audio segments by means of a hard fade;

Fig. 3e shows the concatenation according to the inventive method by means of an example of two inventive audio segments, each of which containing a sound/phone with downstream co-articulation bands, with the last audio segment coming from the end of a phone sequence;

Fig. 3eI shows the concatenation of these audio segments by means of a cross fade;

Fig. 3eII shows the concatenation of these audio segments by means of a hard fade;

Fig. 4 is a schematic representation of the steps of the inventive method for the generation of synthesised acoustical data.

The reference numerals used in the following refer to Fig. 1a and the numbers of the various steps of the method used in the following refer to Fig. 4.

In order to convert for example a text to synthesised speech by means of the invention, it is necessary to divide this text in a preparatory step into a sequence of phonetic characters or phonema, respectively. Preferably, prosodic information corresponding to the text is to be generated as well. The sound or phone sequence, respectively, as well as the prosodic and additional information serve as input values for the inventive method or the inventive device, respectively.

The sounds/phones to be synthesised are supplied to an input unit 101 of the device 1 for the generation of synthesised voice data and stored in a first memory unit 103 (see Fig. 1a). By means of a selection means 105 audio segments are selected from an inventory including audio segments (elements) which is stored in a database 107, or by an upstream synthesis means 108 (which is not part of the invention), which reproduce sounds or phones, respectively, or portions of sounds or phones, respectively, which correspond to the individually input phonetic characters or phonema, respectively, or portions of same and stored in a second memory unit 109 in an order corresponding to the order to the input phonetic characters or phonema, respectively. If the inventory includes portions of phone sequences or of audio segments, the selection unit 105 preferably selects those audio segments which reproduce the highest number of portions of the phone sequences or polyphones, respectively, which correspond to a sequence of phonetic characters or phonema, respectively, from

the input phone sequence or phoneme sequence, respectively, so that a minimum number of audio segments is required for the synthesis of the input phoneme sequence.

If the database 107 or the upstream synthesis means 108 provides an inventory with audio segments of different types, the selection means 105 preferably selects the longest audio segment bands which reproduce portions of the sound sequence/phone sequence in order to synthesise the input sound sequence or phone sequence, respectively, and/or a sequence of sounds/phones from a minimum number of audio segment bands. In this context, it is advantageous to use audio segment bands reproducing linked sounds/phones, which reproduce an earlier static sound/phone and a later dynamic sound phone. In this manner, audio segments are generated which, because of the embedded dynamic sounds/phones invariably begin with a static sound/phone. For this reason, the concatenation procedure for such audio segments is simplified and standardised, because only cross fades are required for this.

In order to achieve a co-articulation-specific concatenation of the audio segment bands to be linked, the concatenation instances of two successive audio segment bands are established with the aid of a concatenation means 111 as follows:

- If an audio segment band is to be used for synthesising the start of the input sound sequence/phone sequence (step 1), an audio segment band is to be selected from the inventory, which reproduces the start of a sound sequence/phone sequence and to be linked with a later audio segment band (see Fig. 3c and step 3 in Fig. 4).

- In the concatenation of a second audio segment band with an earlier first audio segment band, a distinction must be made as to whether the second audio segment band starts with the

reproduction of a static sound/phone or a dynamic sound/phone in order to appropriately make the selection of the instance of concatenation (step 6).

- If the second audio segment band starts with a static sound/phone, then the concatenation is carried out in the form of a cross fade, with the instance of concatenation being set in the downstream portion of the first audio segment band and in the upstream portion of the second audio segment band, with the two bands overlapping in the concatenation or at least bordering on one another (see Figs. 3bI, 3cI, 3dI, and 3eI; concatenation by means of cross fade).

- If the second audio segment band starts with a dynamic sound/phone, then the concatenation is carried out in the form of a hard fade, with the instance of concatenation being set immediately after of the downstream portion of the first audio segment band and immediately before the upstream band of the second audio segment band (see Figs. 3bII, 3cII, 3dII, and 3eII; concatenation by means of hard fade).

In this manner, new audio segments can be generated from the originally available audio segment bands, which start with the reproduction of a static sound/phone. This is achieved in that audio segment bands which start with the reproduction of a dynamic sound/phone are linked later with audio segment bands which start with the reproduction of a static sound/phone. Though this increases the number of audio segments or the volume of the inventory, respectively, can, however, be a computational advantage, because fewer individual concatenations are required for the generation of a phone sequence/phoneme sequence, and concatenations have to carried out only in the form of cross fades. Preferably, the new linked audio segments are supplied to the database 107 or another memory unit 113.

A further advantage of this linking of the original audio segment bands to new longer audio segments results if, for example, a sequence of sounds/phones frequently repeats itself in the input sound sequence/phone sequence. It is then possible to utilise one of the new correspondingly linked audio segments, and it is not necessary to carry out another concatenation of the originally available audio segment bands with each occurrence of this sequence of sounds/phones. Preferably, overlapping co-articulation effects, too, are to be covered, or specific co-articulation effects in the form of additional data is to be assigned to the stored linked audio segment, respectively, when storing such linked audio segments.

If an audio segment band is to be used for synthesising the end of the input sound sequence/phone sequence, an audio segment band is to be selected from the inventory, which reproduces an end of a sound sequence/phone sequence, and to be linked with an earlier audio segment band (see Fig. 3e and step 8 in Fig. 4).

The individual audio segments are stored in a coded form in the database 107, with the coded form of the audio segments, apart from the waveform of the respective audio segment, being able to indicate which type of concatenation (e.g. hard fade, linear or exponential cross fade) is to be carried out with which later audio segment band, and at which instance the concatenation takes place with which later audio segment band. Preferably, the coded form of the audio segments also includes information with respect to the prosody, higher-order co-articulations and transfer functions which are used to achieve an additional improvement of the voice quality.

In the selection of the audio segment bands for synthesising the input sound sequence/phone sequence, the audio segment bands selected as the later ones are such that they correspond

to the properties of the respective earlier audio segment bands, i.a. type of concatenation and concatenation instance. After the selection of the audio segment bands, each of which reproducing portions of the sound sequence/phone sequence, from the database 107 or the upstream synthesising means 108, the concatenation of two successive audio segment bands by means of the concatenation means 111 is carried out as follows. The waveform, the type of concatenation, the concatenation instance as well as any additional information, if required, of the first audio segment band and the second audio segment band are loaded from the database of the synthesising means (Fig. 3b and steps 10 and 11). Preferably such audio segment bands are selected in the above mentioned selection of the audio segment bands, which are in agreement with each other with respect to their type and instance of concatenation. In this case, loading of information with respect to type and instance of concatenation of the second audio segment band is no longer necessary.

For the concatenation of the two audio segment bands, the waveform of the first audio segment band in a later band and the waveform of the second audio segment band in an earlier band, each are processed by means of suitable transfer functions, e.g. multiplied by a suitable weighting function (see Fig. 3b, steps 12 and 13). The lengths of the later band of the first audio segment and of the earlier band of the second audio segment result from the type of concatenation and the time position of the concatenation instance, with these lengths also being able to be stored in the coded form of the audio segments in the database.

If the two audio segment bands are to be linked by means of a cross fade, they are added in an overlapping manner according to the respective instance of concatenation (see Figs. 3bI, 3cI, 3dI, and 3eI; step 15). Preferably, a linear symmetrical

cross fade is to be used herein, however, any other type of cross fade or any type of transfer function can be employed as well. If a concatenation in the form of a hard fade is to be carried out, the two audio segment bands are not joined consecutively in an overlapping manner (see Figs. 3bII, 3cII, 3dII, and 3eII; step 15). As can be seen in Fig. 3bII, the two audio segment bands are arranged immediately successive in time. In order to be able to further process the voice generated in this manner, it is preferably stored in a third memory unit 115.

For the further linking with successive audio segment bands, the audio segments bands linked so far are considered as a first audio segment band (step 16), and the above described linking process is repeated until the entire sound sequence/phone sequence has been synthesised.

For an improvement of the quality of the synthesised voice data, the prosodic and additional information which are input in addition to the sound sequence/phone sequence, are preferably to be considered in the linking of the audio segment bands. By means of known methods, the frequency, duration, amplitude, and/or spectral properties of the audio segment bands can be modified before and/or after the concatenation in such a manner that the synthesised voice data comprises a natural word and/or phrase melody (steps 14, 17, or 18). In this context it is to be preferred to select concatenation instances at places of the audio segment bands, at which they agree in one or several suitable properties.

In order to optimise the transitions between two successive audio segment bands, the processing of the two audio segment bands by means of suitable functions in the area of the concatenation instance is additionally provided, in order to i.a. tune the frequencies, durations, amplitudes, and spectral

properties. The invention additionally permits to take into consideration higher-order acoustical phenomena of a real speech, such as for example higher-order co-articulation effects of style of speech (i.a. whispering, stress, singing voice, falsetto, emotional expression) in the synthesising of the sound sequence/phone sequence. For this purpose, information relating to such higher-order phenomena, is additionally stored in a coded form with the corresponding audio segment bands in order to select only such audio segment bands in the selection which correspond to the higher-order co-articulation properties of the earlier and/or later audio segment bands.

The synthesised voice data generated in this manner preferably have a form which, with the aid of an output means 117, allows to convert the voice data to acoustical voice signals and to store the voice data and/or voice signals in an acoustical, optical, magnetic, or electrical data carrier (step 19).

Generally, inventory elements are generated via the recording of actually spoken speech. Depending on the level of training of the inventory-building speaker, i.e. his or her capability for controlling the speech to be recorded (e.g. to control the tone level of the speech or to speak exactly on one tone level), it is possible to generate identical or similar inventory elements which have displaced boundaries between the solo articulation bands and the co-articulation bands. This results in considerably more possibilities of setting the concatenation points in different places. As a consequence, the quality of a speech to be synthesised can be considerably enhanced.

This invention allows for the first time to generate synthesised voice signals by means of a co-articulation-specific concatenation of individual audio segment bands, because the instance of concatenation is selected depending on the res-

pective audio segment bands to be linked. In this manner, a synthesised speech can be generated which is no longer distinguishable from a naturally spoken speech. Contrary to known methods or devices, the audio segments used herein are not generated by speaking or recording, respectively, complete words, in order to ensure an authentic voice quality. It is therefore possible by means of this invention to generate synthesised speech of any contents with the quality of an actually spoken speech.

Although this invention is described by way of the example of the speech synthesis, it is not limited to the field of synthesised speech, but can be used for synthesising any acoustical data or any sound events, respectively. This invention can therefore be employed for the generation and/or provision of synthesised voice data and/or voice signals for any language or dialect, as well as for the synthesis of music.

Claims

1. A method for the co-articulation-specific concatenation of audio segments, in order to generate synthesised acoustical data which reproduces a sequence of concatenated sounds/phones, comprising the following steps:
- selecting at least two audio segments which contain bands, each of which reproducing a portion of a sound/phone or a portion of a sound/phone sequence,
- establishing a band to be used of an earlier audio segment;
- establishing a band to be used of a later audio segment, which begins with the later audio segment and ends with the co-articulation band of the later audio segment which follows the initially used solo articulation band;
- with the duration and position of the bands to be used being determined as a function of the earlier and later audio segments; and
- concatenating the established band of the earlier audio segment with the established band of the later audio segment, in that the instance of concatenation, as a function of properties of the used band of the later audio segment, is set in a band which begins immediately before the used band of the later audio segment and ends with same.

2. The method according to Claim 1, characterised in that
- the instance of concatenation is set in a band which lies in the vicinity of the boundaries of the initially to be used solo articulation band of the later audio segment, if the band of same to be used reproduces a static sound/phone at the beginning; and
- a downstream portion of the band to be used of the earlier audio segment and an upstream portion of the band to be used of the later audio segment are processed by means of suitable transfer functions and added in an overlapping manner (cross

fade), with the transfer functions and the length of an overlapping portion of the two bands being determined depending on the audio segments to be concatenated.

3. The method according to Claim 1 or 2, characterised in that
- the instance of concatenation is set in a band which lies immediately before the band to be used of the later audio segment, if the used band of same reproduces a dynamic sound/phone at the beginning; and
- a downstream portion of the band to be used of the earlier audio segment and an upstream portion of the band to be used of the later audio segment are processed by means of suitable transfer functions and joined in a non-overlapping manner (hard fade), with the transfer functions being determined depending on the acoustical data to be synthesised.

4. The method according to one of Claims 1 to 3, characterised in that for a sound/phone or a portion of the sequence of concatenated sounds/phones at the start of the concatenated sound/phone sequence a band of an audio segment is selected so that the start of the band reproduces the properties of the start of the concatenated sound/phone sequence.

5. The method according to one of Claims 1 to 4, characterised in that for a sound/phone or a portion of the sequence of concatenated sounds/phones at the end of the concatenated sound/phone sequence a band of an audio segment is selected so that the end of the band reproduces the properties of the end of the concatenated sound/phone sequence.

6. The method according to one of Claims 1 to 5, characterised in that the voice data to the synthesised is combined in groups, each of which being described by an individual audio segment.

7. The method according to one of Claims 1 to 6, characterised in that an audio segment is selected for the later audio segment band, which reproduces the highest number of successive portions of the sounds/phones of the sound/phone sequence, in order to use the smallest number of audio segment bands in the generation of the synthesised acoustical data.

8. The method according to one of Claims 1 to 7, characterised in that a processing of the used bands of individual audio segments is carried out by means of suitable functions depending on properties of the concatenated sound/phone sequence, with these properties involving i.a. a modification of the frequency, the duration, the amplitude, or the spectrum.

9. The method according to one of Claims 1 to 8, characterised in that a processing of the used bands of individual audio segments is carried out by means of suitable functions in a band, in which the instance of concatenation lies, with these functions involving i.a. a modification of the frequency, the duration, the amplitude, or the spectrum.

10. The method according to one of Claims 1 to 9, characterised in that the instance of concatenation is set in places of the bands to be used of the earlier and/or later audio segment, in which the two used bands are in agreement with respect to one or several suitable properties, with these properties including i.a.: zero point, amplitude values, gradients, derivatives of any degree, spectra, tone levels, amplitude values within a frequency band, volume, style of speech, emotion of speech, or other properties covered in the phone classification scheme.

11. The method according to one of Claims 1 to 10, characterised in that

- the selection of the used bands of individual audio segments, their processing, their variation, as well as their concatenation are additionally carried out with the application of heuristic knowledge which is obtained by an additionally carried out heuristic method.

12. The method according to one of Claims 1 to 11, characterised in that
- the acoustical data to be synthesised is voice data, and the sounds are phones.

13. The method according to one of Claims 2 to 12, characterised in that
- the static phones include vowels, diphtongs, liquids, vibrants, fricativen and nasals.

14. The method according to one of Claims 3 to 13, characterised in that and
- the dynamic phones include plosives, affricates, glottal stops, and click sounds.

15. The method according to one of Claims 1 to 14, characterised in that
- a conversion of the synthesised acoustical data to acoustical signals and/or voice signals is carried out.

16. A device for the co-articulation-specific concatenation of audio segments, in order to generate synthesised acoustical data which reproduces a sequence of phones, comprising:
- a database (107) in which audio segments are stored, each of which reproducing portion of a phone or portions of a sequence of (concatenated) phones;
- and/or any upstream synthesis means (108) which supplies audio segments;

- a means (105) for the selection of at least two audio segments from the database (107) and/or the upstream synthesis means (108); and
- a means (111) for the concatenation of audio segments,
characterised in that the concatenation means (111) is suited for
- defining a band to be used of an earlier audio segment;
- defining a portion to be used of a later audio segment in a band which starts with the later audio segment and ends after a co-articulation band of the later audio segment, which follows after the initially used solo articulation band;
- determining the duration and position of the used bands depending on the earlier and later audio segments; and
- concatenating the used band of the earlier audio segment with the used band of the later audio segment by defining the instance of concatenation as a function of properties of the used band of the later audio segment in a band which starts immediately before the used band of the later audio segment and ends with same.

17. The device according to Claim 16, characterised in that the concatenation means (111) comprises:
- means for the concatenation of the used band of the earlier audio segment with the used band of the later audio segment, whose used band reproduces a static phone at the beginning in the vicinity of the boundaries of the initially occurring solo articulation band of the used band of the later audio segment;
- means for processing a downstream portion of the used band of the earlier audio segment and an upstream portion of the used band of the later audio segment by suitable transfer functions; and
- means for the overlapping addition of the two bands in an overlapping portion (cross fade), which depends on the audio segments to be concatenated, with the transfer functions and

the length of an overlapping portion of the two bands being determined depending on the acoustical data to be synthesised.

18. The device according to Claim 16 or 17, characterised in that the concatenation (111) means comprises:
- means for the concatenation of the used band of the earlier audio segment with the used band of the later audio segment, whose used band reproduces a dynamic phone at the beginning, immediately before the used band of the later audio segment;
- means for processing a downstream portion of the used band of the earlier audio segment and an upstream portion of the used band of the later audio segment by suitable transfer functions, with the transfer functions being determined depending on the acoustical data to be synthesised; and
- means for the non-overlapping joining of the two audio segments.

19. The device according to one of Claims 16 to 18, characterised in that the database (107) includes audio segments or the upstream synthesis means (108) supplies audio segments which comprise bands which at the start reproduce a phone or a portion of the concatenated phone sequence at the start of the concatenated phone sequence.

20. The device according to one of Claims 16 to 19, characterised in that the database (107) includes audio segments or the upstream synthesis means (108) supplies audio segments which comprise bands, whose ends reproduce a phone or a portion of the concatenated phone sequence at the end of the concatenated phone sequence.

21. The device according to one of Claims 16 to 19, characterised in that the database (107) includes a group of audio segments or the upstream synthesis means (108) supplies audio

segments which comprise bands, whose starts each reproduce only a static phone.

22. The device according to one of Claims 16 to 21, characterised in that the concatenation means (111) comprises:
- means for the generation of further audio segments by concatenation of audio segments, with the starts of the bands each reproducing a static phone, each with a band of a later audio segment whose used band reproduces a dynamic phone at the start, and
- a means which supplies the further audio segments to the database (107) or the selection means (105).

23. The device according to one of Claims 16 to 22, characterised in that, in the selection of the audio segment bands from the database (107) or the upstream synthesis means (108), the selection means (105) is suited to select the audio segments which reproduce the greatest number of successive portions of concatenated phones of the concatenated phone sequence.

24. The device according to one of Claims 16 to 23, characterised in that the concatenation means (111) comprises means for processing the used bands of individual audio segments with the aid of suitable functions, depending on properties of the concatenated phone sequence, with the functions involving among others a modification of the frequency, the duration, the amplitude, or the spectrum.

25. The device according to one of Claims 16 to 24, characterised in that
- the concatenation means (111) comprises means for processing the used bands of individual audio segments with the aid of suitable functions in a band including the instance of conca-

tenation, with this function involving i.a. a modification of the frequency, the duration, the amplitude, or the spectrum.

26. The device according to one of Claims 16 to 25, characterised in that
- the concatenation means (111) comprises means for the selection of the instance of concatenation in a place in the used bands of the earlier and/or the later audio segment, in which the two used bands are in agreement with respect to one or several suitable properties, with these properties including i.a.: zero points, amplitude values, gradients, derivatives of any degree, spectra, tone levels, amplitude values in a frequency band, volume, style of speech, emotion of speech, or other properties covered in the phone classification scheme.

27. The device according to one of Claims 16 to 26, characterised in that
- the selection means (105) comprises means for the implementation of heuristic knowledge which relates to the selection of the used bands of the individual audio segments, their processing, their variation, as well as their concatenation.

28. The device according to one of Claims 16 to 27, characterised in that
- the database (107) includes audio segments or the upstream synthesis means (108) supplies audio segments which include bands, each of which reproducing at least a portion of a sound or phone, respectively, a sound or phone, respectively, portions of phone sequences or polyphones, respectively, or sound sequences or polyphones, respectively.

29. The device according to one of Claims 17 to 28, characterised in that
the data base (107) includes audio segments or the upstream synthesis means (108) supplies audio segments, with a static

sound corresponding to a static phone and comprising vowels, diphtongs, liquids, vibrants, fricatives, and nasals.

30. The device according to one of Claims 18 to 29, characterised in that
- the database (107) includes audio segments or the upstream synthesis means (108) supplies audio segments, with a dynamic sound corresponding to a dynamic phone and comprising plosives, affricates, glottal stops, and klick speech.

31. The device according to one of Claims 16 to 30, characterised in that
- the concatenation means (111) is suitable to generate synthesised voice data by means of the concatenation of audio segments.

32. The device according to one of Claims 16 to 31, characterised in that
- means (117) are provided for the conversion of the synthesised acoustical data to acoustical signals and/or voice signals.

33. A data carrier which includes a computer program for the co-articulation-specific concatenation of audio segments in order to generate synthesised acoustical data which reproduces a sequence of concatenated phones, comprising the following steps:
- selection of at least two audio segments which contain bands, each of which reproducing a portion of a sound/phone or a portion of a sound/phone sequence,
characterised by the steps of:
- establishing a band to be used of an earlier audio segment;
- establishing a band to be used of a later audio segment, which begins with the later audio segment and ends with the

co-articulation band of the later audio segment which follows the initially used solo articulation band;
- with the duration and position of the bands to be used being determined as a function of the earlier and later audio segments; and
- concatenating the established band of the earlier audio segment with the established band of the later audio segment, in that the instance of concatenation, as a function of properties of the used band of the later audio segment, is set in its established band which starts immediately before the band to be used of the later audio segment and ends with same.

34. The data carrier according to Claim 33, characterised in that the computer program selects the instance of the concatenation of the used band of the second audio segment with the used band of the first audio segment in such a manner that
- the instance of concatenation is set in a band which lies in the vicinity of the boundaries of the initially used solo articulation band of the later audio segment, if its used band reproduces a static phone at the start;
- a downstream portion of the used band of the earlier audio segment and an upstream portion of the used band of the later audio segment are processed by suitable transfer functions and added in an overlapping manner (cross fade), with the transfer functions and the length of an overlapping portion of the two bands being determined depending on the audio segments to be concatenated.

35. The data carrier according to Claim 33 or 34, characterised in that the computer program selects the instance of the concatenation of the used band of the second audio segment with the used band of the first audio segment in such a manner that

- the instance of concatenation is set in a band which lies immediately before the used band of the later audio segment, if its used band reproduces a dynamic phone at the start;
- a downstream portion of the used band of the earlier audio segment and an upstream portion of the used band of the later audio segment are processed by suitable transfer functions and added in a non-overlapping manner (hard fade), with the transfer functions being determined depending on the audio segments to be concatenated.

36. The data carrier according to one of Claims 33 to 35, characterised in that the computer program selects a band of an audio segment for a phone or a portion of the sequence of concatenated phones at the start of the concatenated phone sequence, the start of which reproduces the properties of the start of the concatenated sequence of phones.

37. The data carrier according to one of Claims 33 to 36, characterised in that the computer program selects a band of an audio segment for a phone or a portion of the sequence of concatenated phones at the end of the concatenated phone sequence, the end of which reproduces the properties of the end of the concatenated sequence of phones.

38. The data carrier according to one of Claims 33 to 37, characterised in that the computer program carries out a processing of the used bands of individual audio segments with the aid of suitable functions depending on properties of the phone sequence, with the functions involving i.a. modification of the frequency, the duration, the amplitude, or the spectrum.

39. The data carrier according to one of Claims 33 to 38, characterised in that the computer program selects an audio segment band for the later audio segment band which reproduces

the highest number of successive portions of the concatenated phones in the phone sequence, in order to use the smallest number of audio segment bands in the generation of the synthesised acoustical data.

40. The data carrier according to one of Claims 39 to 45, characterised in that the computer program carries out a processing of the used bands of individual audio segments with the aid of suitable functions in a band in which the instance of concatenation lies, with these functions involving i.a. a modification of the frequency, the duration, the amplitude, or the spectrum.

41. The data carrier according to one of Claims 33 to 40, characterised in that the computer program establishes the instance of concatenation in a place of the used bands of the first and/or the second audio segment, in which the two used bands are in agreement with respect to one or several suitable properties, with these properties including i.a.: zero points, amplitude values, gradients, derivatives of any degree, spectra, tone levels, amplitude values in a frequency band, volume, style of speech, emotion of speech, or other properties covered in the phone classification scheme.

42. The data carrier according to one of Claims 33 to 41, characterised in that the computer program carries out an implementation of heuristic knowledge which relates to the selection of the used bands of the individual audio segments, their processing, their variation, as well as their concatenation.

43. The data carrier according to one of Claims 33 to 42, characterised in that the computer program is suited for the generation of synthesised voice data, with the sounds being phones.

44. The data carrier according to one of Claims 34 to 42, characterised in that the computer program is suited for the generation of static phones, with the static phones comprising vowels, diphtongs, liquids, vibrants, fricatives, and nasals.

45. The data carrier according to one of Claims 35 to 44, characterised in that the computer program is suited for the generation of dynamic phones, with the dynamic phones comprising plosives, affricates, glottal stops, and klick speech.

46. The data carrier according to one of Claims 33 to 45, characterised in that the computer program converts the synthesised acoustical data to acoustical convertible data and/or voice signals.

47. Synthesised voice signals which consist of a sequence of sounds or phones, respectively, with the voice signals being generated in that:
- at least two audio segments are selected which reproduce the sounds or phones, respectively; and
- the audio segments are linked by a co-articulation-specific concatenation, with
- one band to be used of an earlier audio segment being established;
- one band to be used of a later audio segment being established which starts with the later audio segment and ends with the co-articulation band of the later audio segment, following the initially used solo articulation band;
- with the duration and position of the bands to be used being determined depending on the audio segments; and
- the used bands of the audio segments being concatenated in a co-articulation-specific manner, in that the instance of concatenation, as a function of properties of the used band of the later audio segment, is set in a band which starts imme-

diately before the used band of the later audio segment and ends with same.

48. The synthesised voice signals according to Claim 47, characterised in that the voice signals are generated in that
- the audio segments are concatenated in an instance which lies in the vicinity of the boundaries of the later audio segment, if the start of this band reproduces a static sound or phone, respectively, with the static phone being a vowel, a diphtong, a liquid, a fricative, a vibrant, or a nasal; and
- a downstream portion of the used band of the earlier audio segment and an upstream portion of the used band of the later audio segment are processed by means of suitable transfer function and both bands are added in an overlapping manner (cross fade), with the transfer functions and the length of an overlapping portion of the two bands being determined depending on the audio segments to be concatenated.

49. The synthesised voice signals according to Claim 47 or 48, characterised in that the voice signals are generated in that
- the audio segments are concatenated in an instance which lies immediately before the used band of the later audio segment, if the start of this band reproduces a dynamic sound or phone, respectively, with the dynamic phone being a plosive, an affricate, a glottal stop, or klick speech; and
- a downstream portion of the used band of the earlier audio segment and an upstream portion of the used band of the later audio segment are processed by means of suitable transfer functions and both bands are joined in a non-overlapping manner (hard fade), with the transfer functions being determined depending on the audio segments to be concatenated.

50. The synthesised voice signals according to one of Claims 47 to 49, characterised in that

- the first sound or the first phone, respectively, or a portion of the first phone sequence or of the first polyphone, respectively, in the sequence is generated by an audio segment, whose used band at the start reproduces the properties of the start of the sequence.

51. The synthesised voice signals according to one of Claims 47 to 50, characterised in that
- the last sound or the last phone, respectively, or a portion of the last phone sequence or of the last polyphone, respectively, in the sequence is generated by an audio segment, whose used band at the end reproduces the properties of the end of the sequence.

52. The synthesised voice signals according to one of Claims 47 to 51, characterised in that
- the voice signals are generated in that later bands of audio segments, beginning with the reproduction of a dynamic sound or phone, respectively, are concatenated with earlier bands of audio segments, beginning with the reproduction of a static sound or phone, respectively.

53. The synthesised voice signals according to one of Claims 47 to 52, characterised in that
- such audio segments are selected which reproduce the highest number of portions of sounds or phones, respectively, of the sequence, in order to use the smallest number of audio segment bands in the generation of the voice signals.

54. The synthesised voice signals according to one of Claims 47 to 53, characterised in that
- the voice signals are generated by the concatenation of the used bands of audio segments which are processed with the aid of suitable functions depending on properties of the sound sequence or phone sequence, respectively, with the functions in-

volving i.a. a modification of the frequency, the duration, the amplitude, or the spectrum.

55. The synthesised voice signals according to one of Claims 47 to 54, characterised in that
- the voice signals are generated by the concatenation of the used bands of audio segments which are processed with the aid of suitable functions depending on properties of the sound sequence or phone sequence, respectively, in an area in which the instance of concatenation lies, with these properties including i.a. a modification of the frequency, the duration, the amplitude, or the spectrum.

56. The synthesised voice signals according to one of Claims 47 to 55, characterised in that the instance of concatenation lies at a place in the used bands of the earlier and/or the later audio segment, in which the two used bands are in agreement with respect to one or several suitable properties, with these properties including i.a.: zero points, amplitude values, gradients, derivatives of any degree, spectra, tone levels, amplitude values in a frequency band, volume, style of speech, emotion of speech, or other properties covered in the phone classification scheme.

57. The synthesised voice signals according to one of Claims 47 to 56, characterised in that the voice signals are suited for a conversion to acoustic signals.

58. An acoustical, optical, magnetic, or electrical data storage which contains audio segments in order generate synthesised acoustical data by means of a concatenation of used bands of the audio segments, utilising the methods according to Claim 1, or the device according to Claim 16, or the data carrier according to Claim 33.

59. The data storage according to Claim 58, characterised in that a group of the audio segments reproduces sounds or phones, respectively, or portions of sounds or phones, respectively.

60. The data storage according to Claim 58 or 59, characterised in that a group of the audio segments reproduces phone sequences or portions of phone sequences or polyphones, respectively, or portions of polyphones.

61. The data storage according to one of Claims 58 to 60, characterised in that a group of audio segments is provided whose used bands start with a static sound or phone, respectively, with the static phones comprising vowels, diphtongs, liquids, fricatives, vibrants, and nasals.

62. The data storage according to one of Claims 58 to 61, characterised in that audio segments are provided which are suitable for the conversion to acoustical signals

63. The data storage according to one of Claims 58 to 62, which additionally contains information in order to carry out a processing of the used bands of individual audio segments with the aid of suitable functions depending on properties of the acoustical data to be synthesised, with the functions involving i.a. a modification of the frequency, the duration, the amplitude, or the spectrum.

64. The data storage according to one of Claims 58 to 63, which additionally contains information relating to a processing of the used bands of individual audio segments with the aid of suitable functions in a band in which the instance of concatenation lies, with this function involving i.a. a modification of the frequency, the duration, the amplitude, or the spectrum.

65. The data storage according to one of Claims 58 to 64, which additionally provides linked audio segments, whose instance of concatenation lies at a place of the used bands of the earlier and/or later audio segment, where both used bands are in agreement with respect to one or several suitable properties with these properties being i.a.: zero points, amplitude values, gradients, derivatives of any degree, spectra, tone levels, amplitude values in a frequency band, volume, style of speech, emotion of speech, or other properties covered in the phone classification scheme.

66. The data storage according to one of Claims 51 to 58, which additionally contains information in the form of heuristic knowledge, which relates to the selection of the used bands of the individual audio segments, their processing, their variation, as well as their concatenation.

67. Sound carrier which contains data which at least partially is synthesised acoustical data which were generated
- by means of the method according to Claim 1, or
- by means of the device according to Claim 16, or
- by utilising the data carrier according to Claim 58, or
- by utilising a data storage according to Claim 58, or
- which are the voice signals according to Claim 47.

68. The sound carrier according to Claim 68, characterised in that the synthesised acoustical data is synthesised voice data.

Abstract

The invention enables the synthesising of any acoustical data by a concatenation of individual audio segment bands, with the instances in which the respective concatenation of two successive audio segment bands take place being established as a function of properties of the audio segments. In this manner, synthesised acoustical data can be generated which, after a conversion to acoustical signals, do not differ from corresponding, naturally generated acoustical signals. In particular, the invention permits the generation of synthesised voice data under consideration of co-articulatory effects by means of a concatenation of individual voice audio segments. The voice data provided in this manner can be converted to voice signals which cannot be distinguished from a naturally spoken language.

Figur 1a:

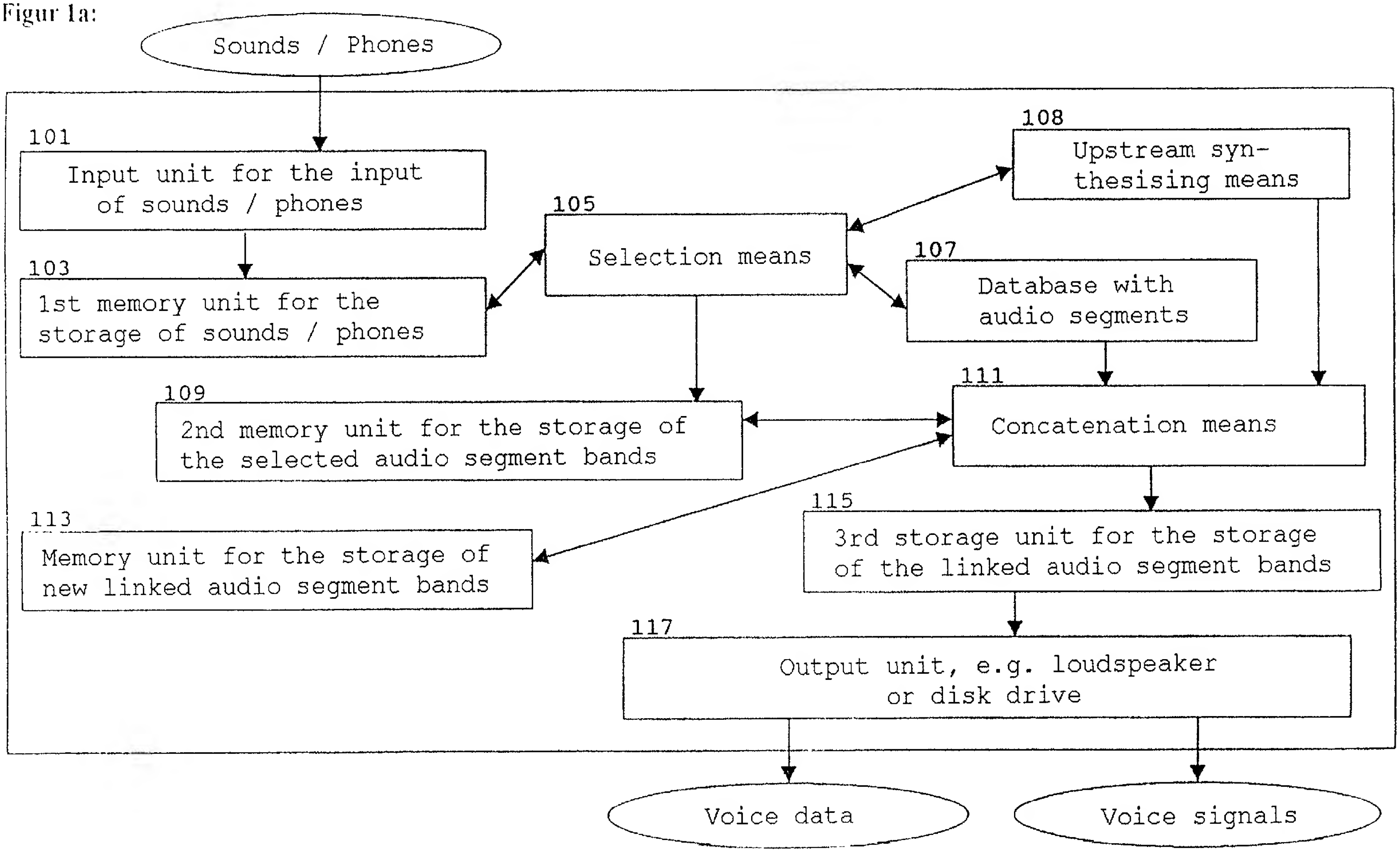
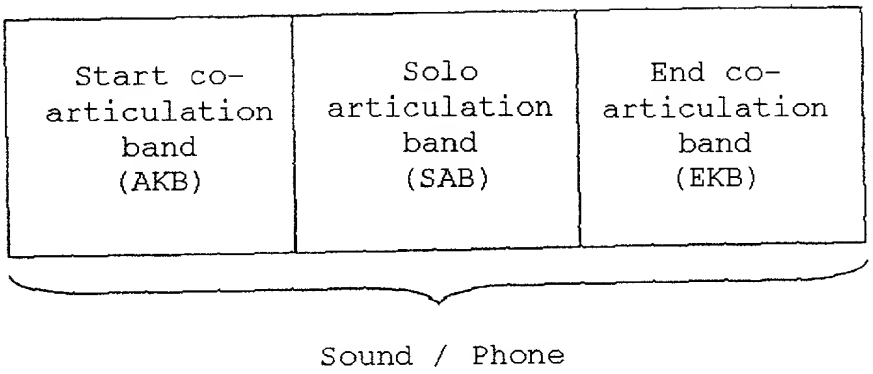

Sounds / Phones
101
Input unit for the input of sounds / phones
103
1st memory unit for the storage of sounds / phones
105
Selection means
108
Upstream synthesising means
107
Database with audio segments
109
2nd memory unit for the storage of the selected audio segment bands
111
Concatenation means
113
Memory unit for the storage of new linked audio segment bands
115
3rd storage unit for the storage of the linked audio segment bands
117
Output unit, e.g. loudspeaker or disk drive
Voice data
Voice signals

Figure 1b: Structure of a sound / phone



Figs. 2a to 2i: Structures of the audio segments

Fig. 2a:

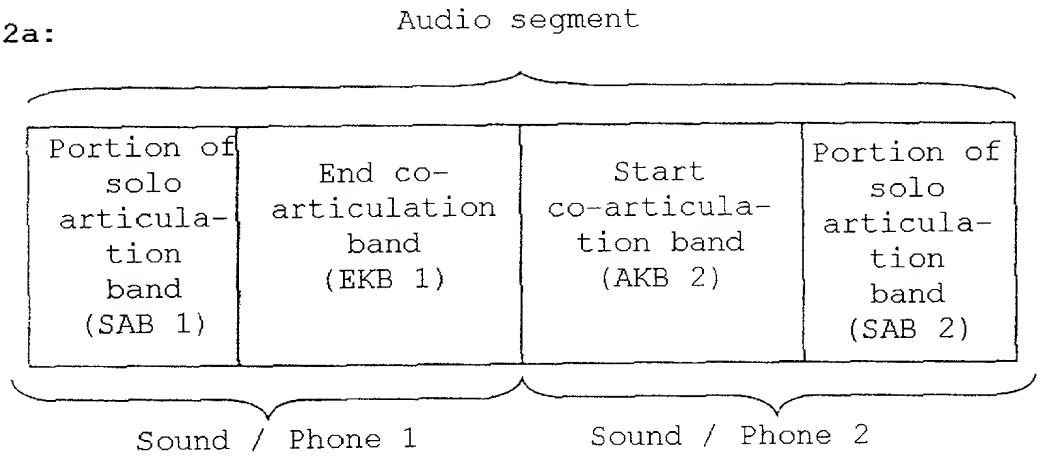


Fig. 2b:

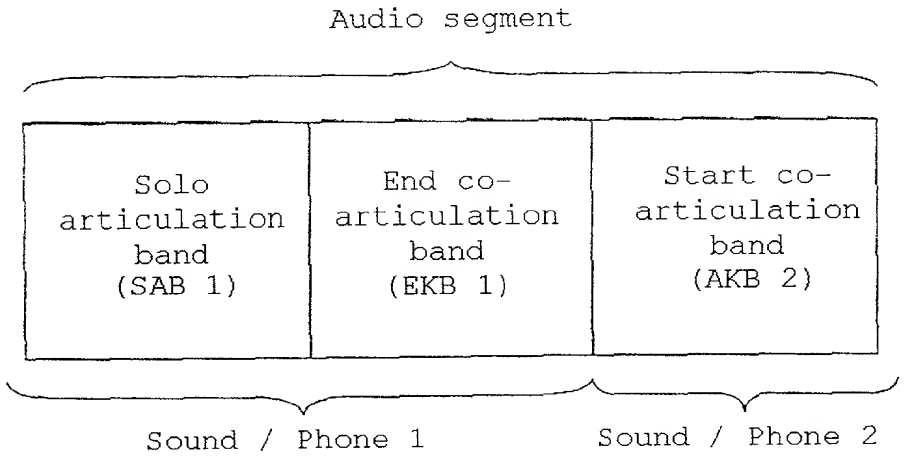


Fig. 2c:

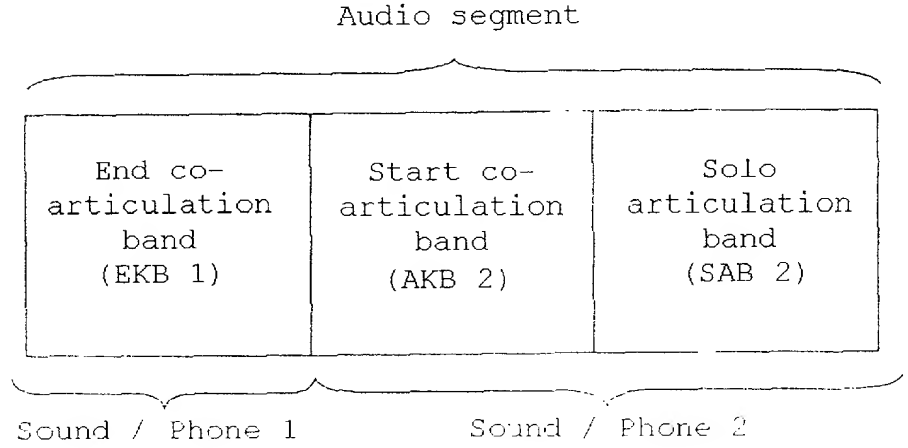


Fig. 2d:

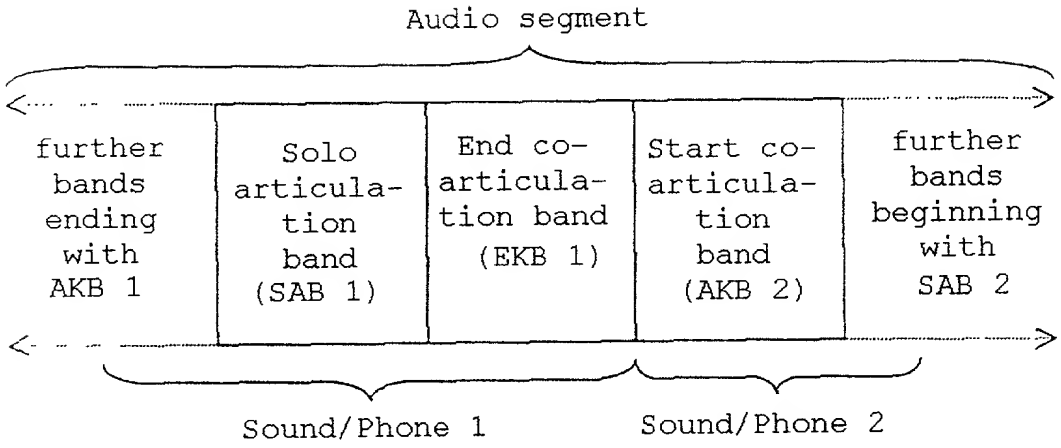


Fig. 2e:

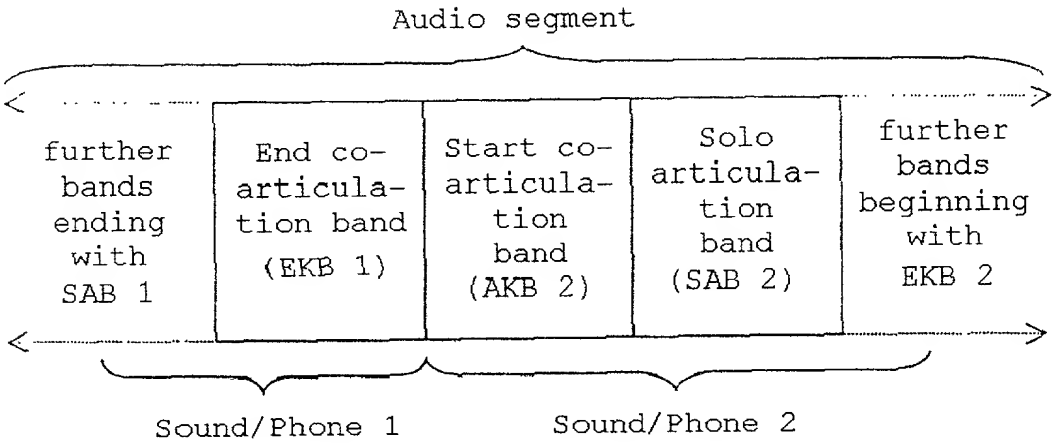


Fig. 2f:

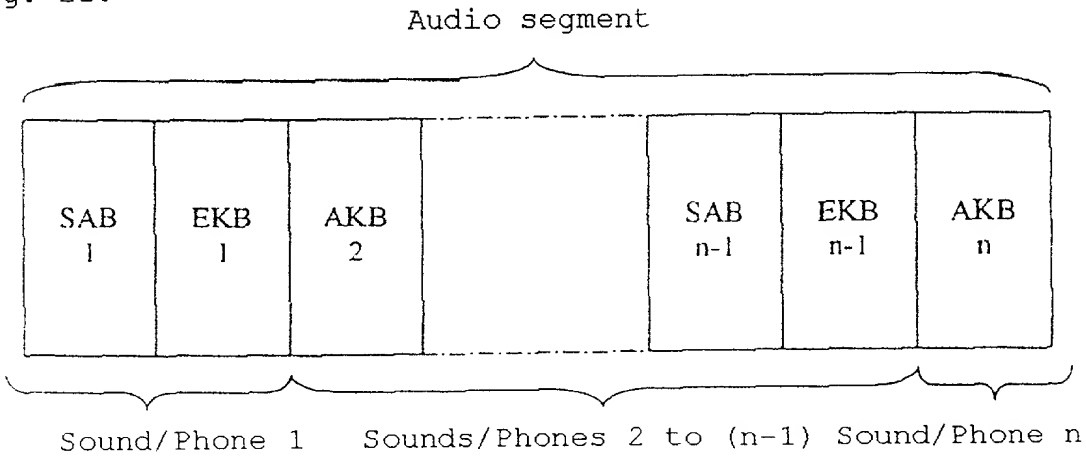


Fig. 2g:

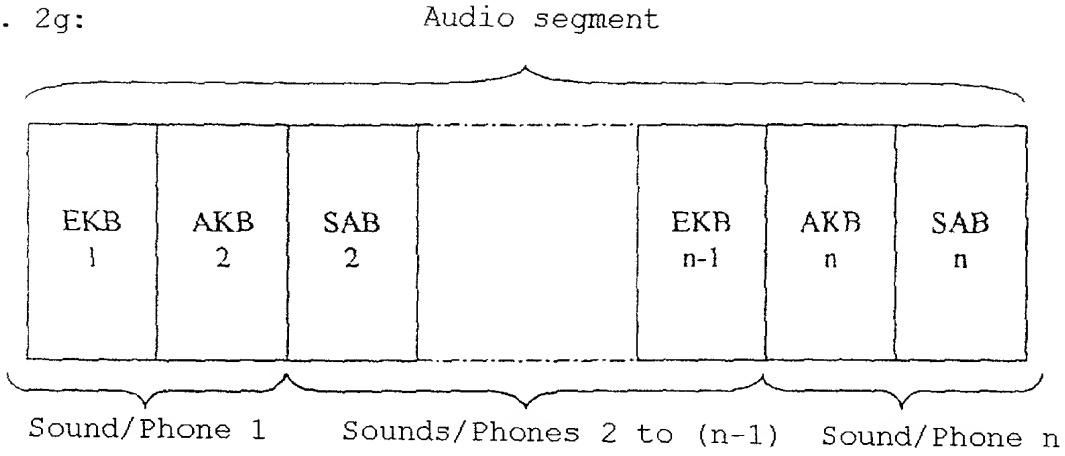


Fig. 2h:

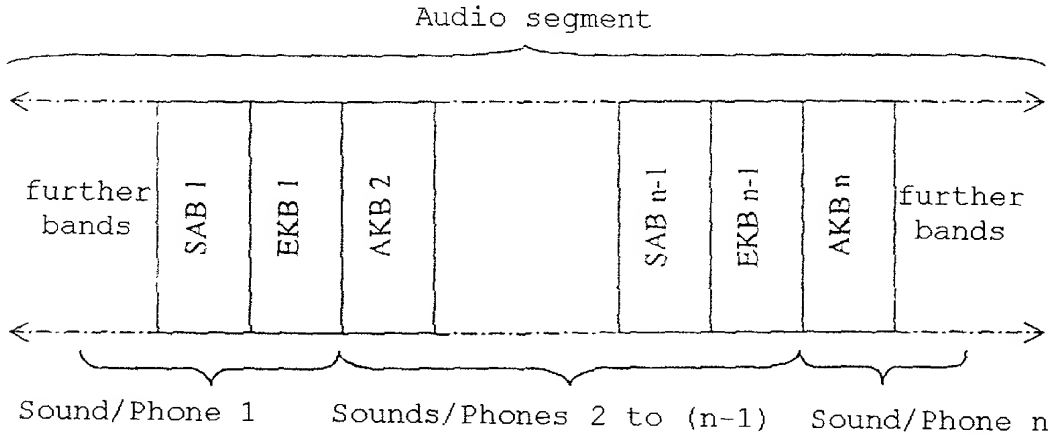


Fig. 2i:

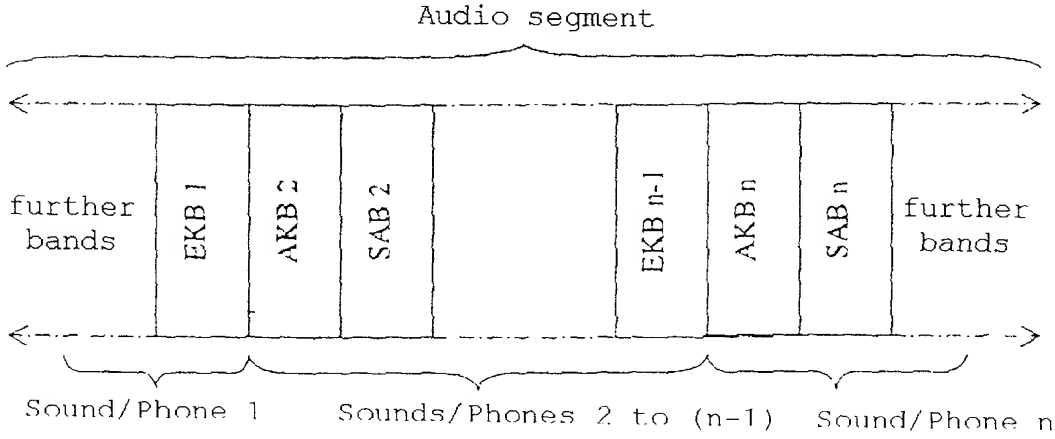


Fig. 2j:

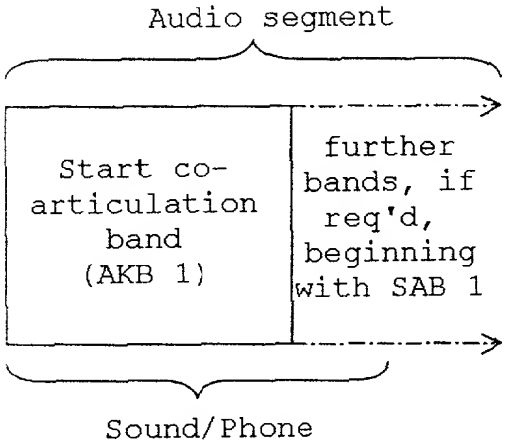


Fig. 2k:

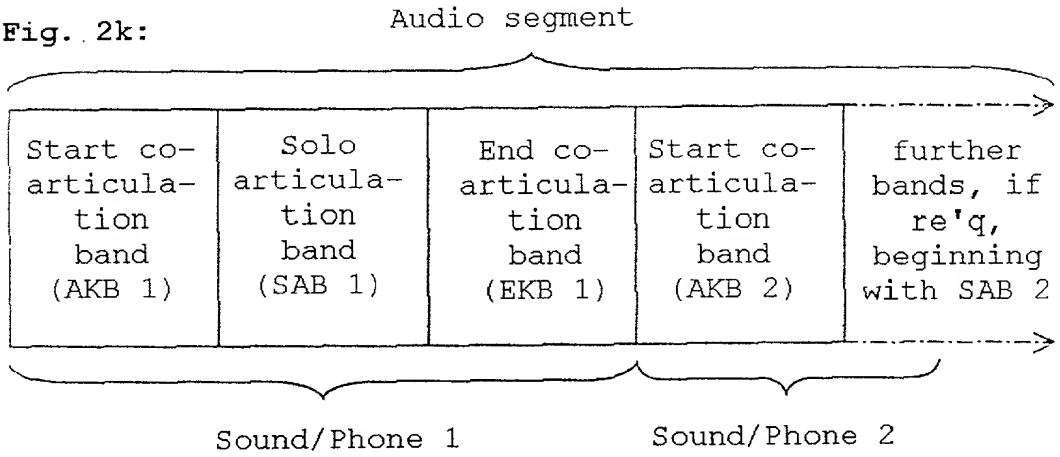
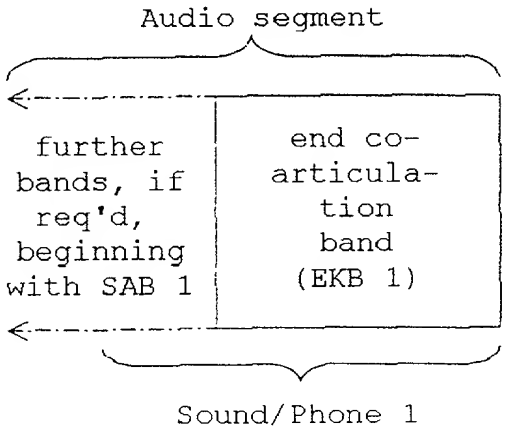


Fig. 2l:



Figs. 3a to 3d: Concatenation

Fig. 3a:

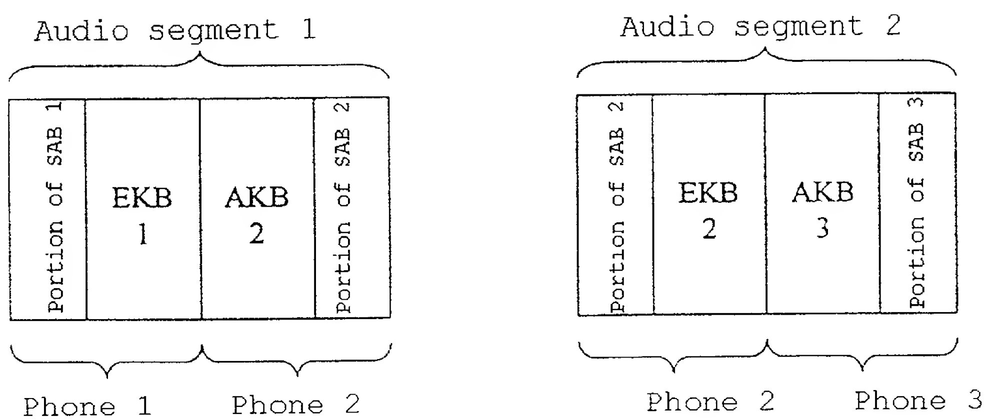


Fig. 3aI:

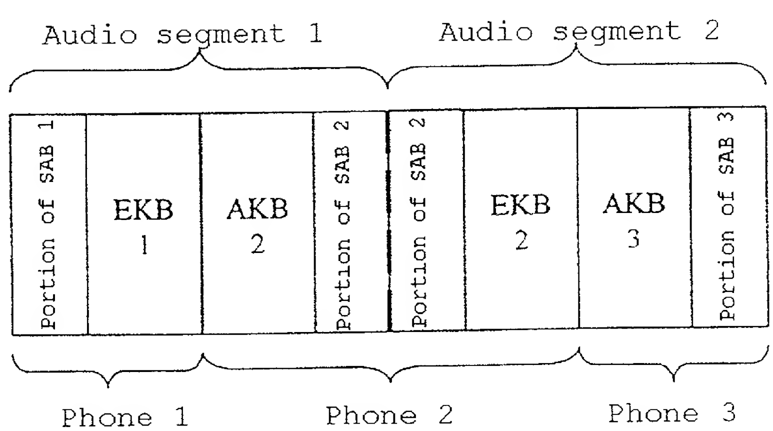


Fig. 3b:

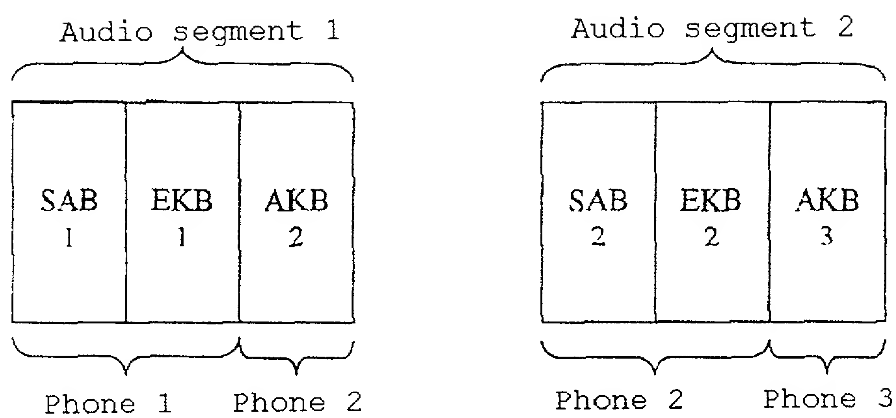


Fig. 3bI:

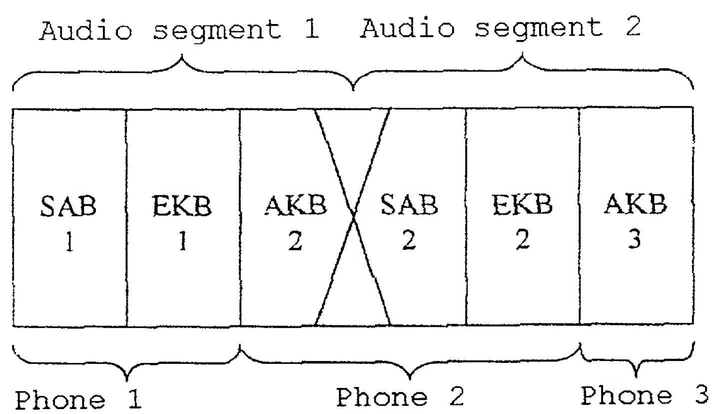


Fig. 3bII:

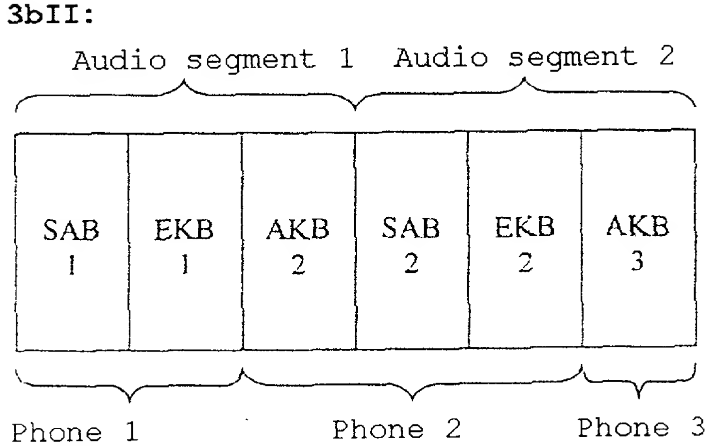


Fig. 3c:

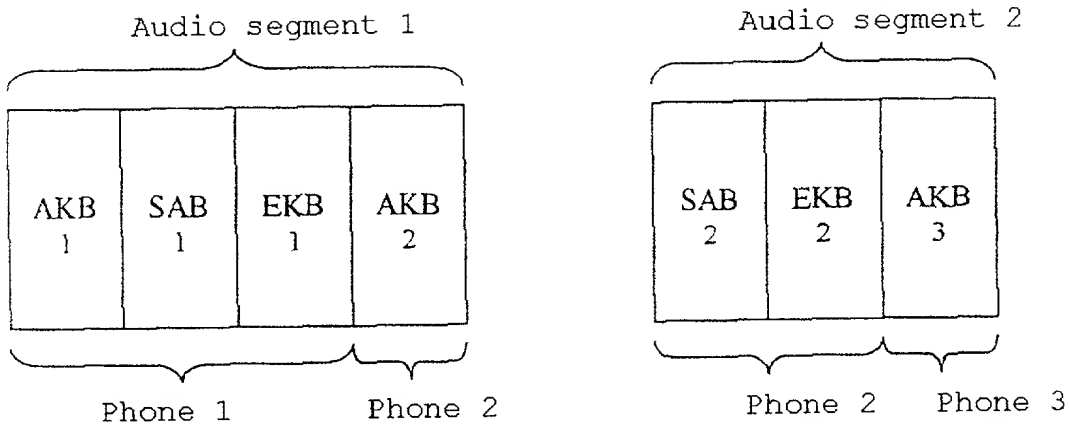


Fig. 3cI:

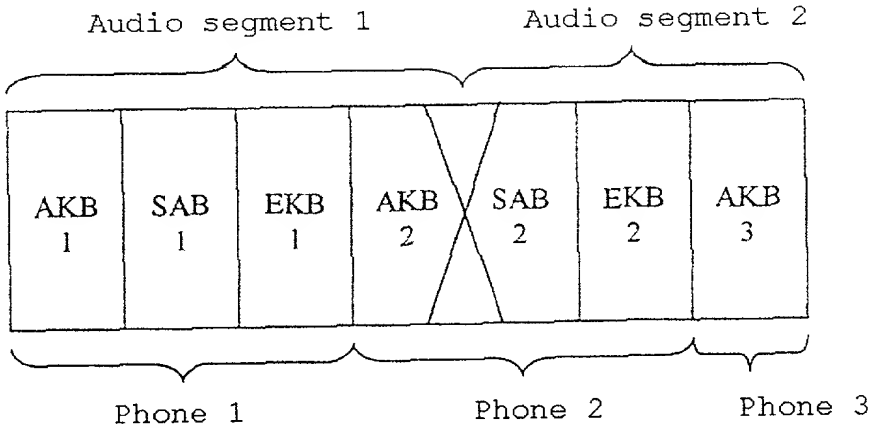


Fig. 3cII:

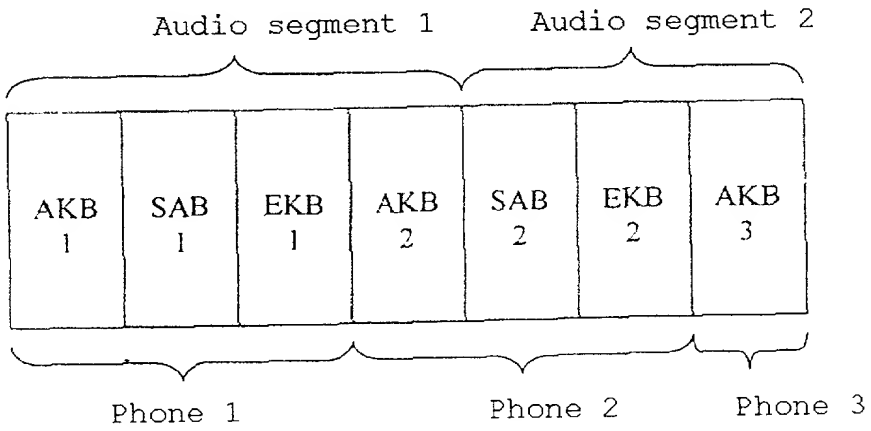


Fig. 3d:

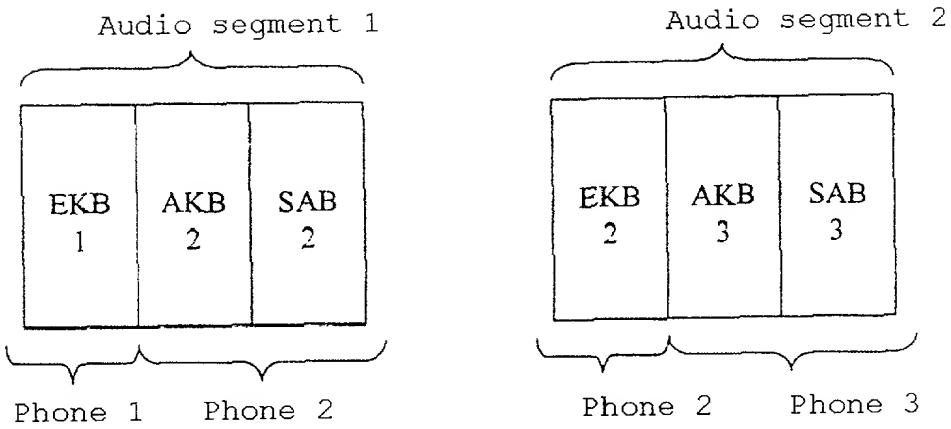


Fig. 3dI:

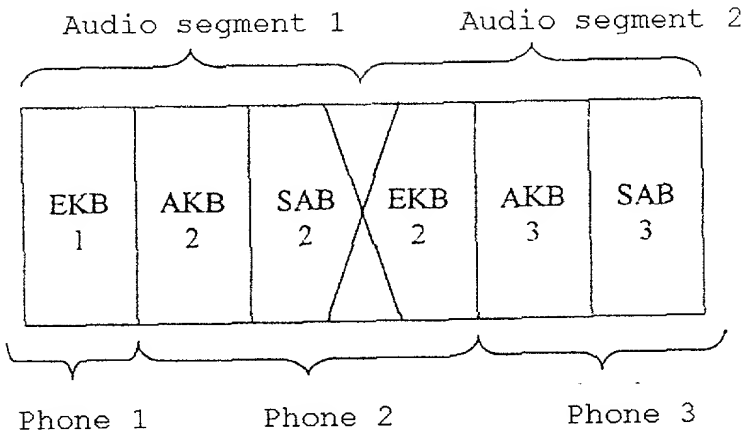


Fig. 3dII:

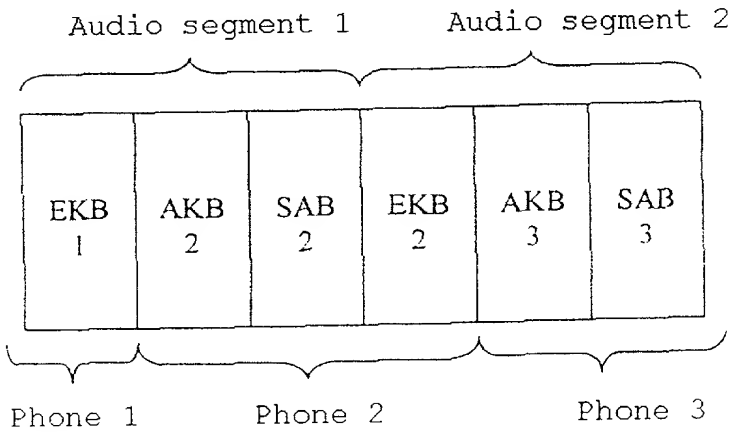


Fig. 3e:

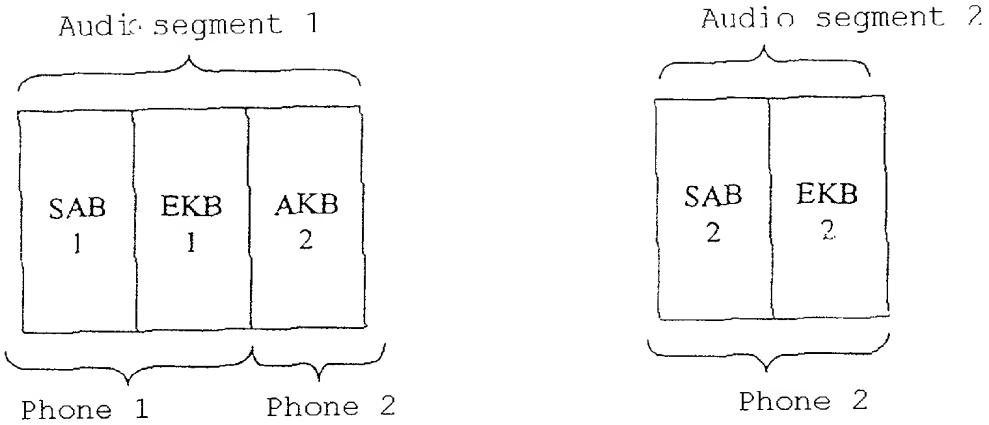


Fig. 3eI:

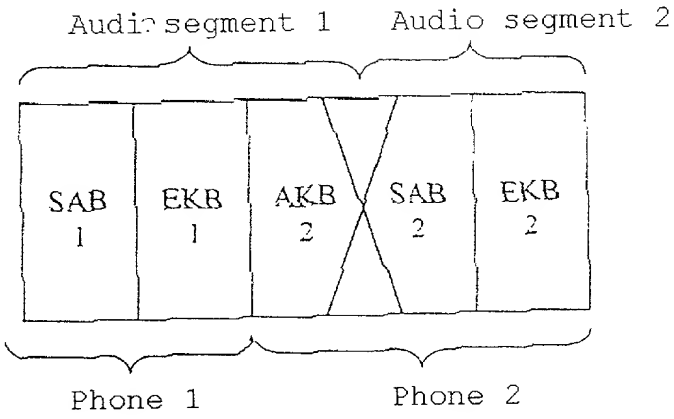


Fig. 3eII:

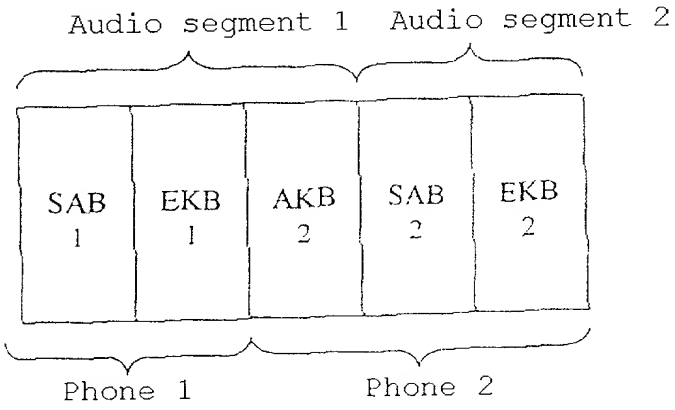


Fig. 4, Part 1

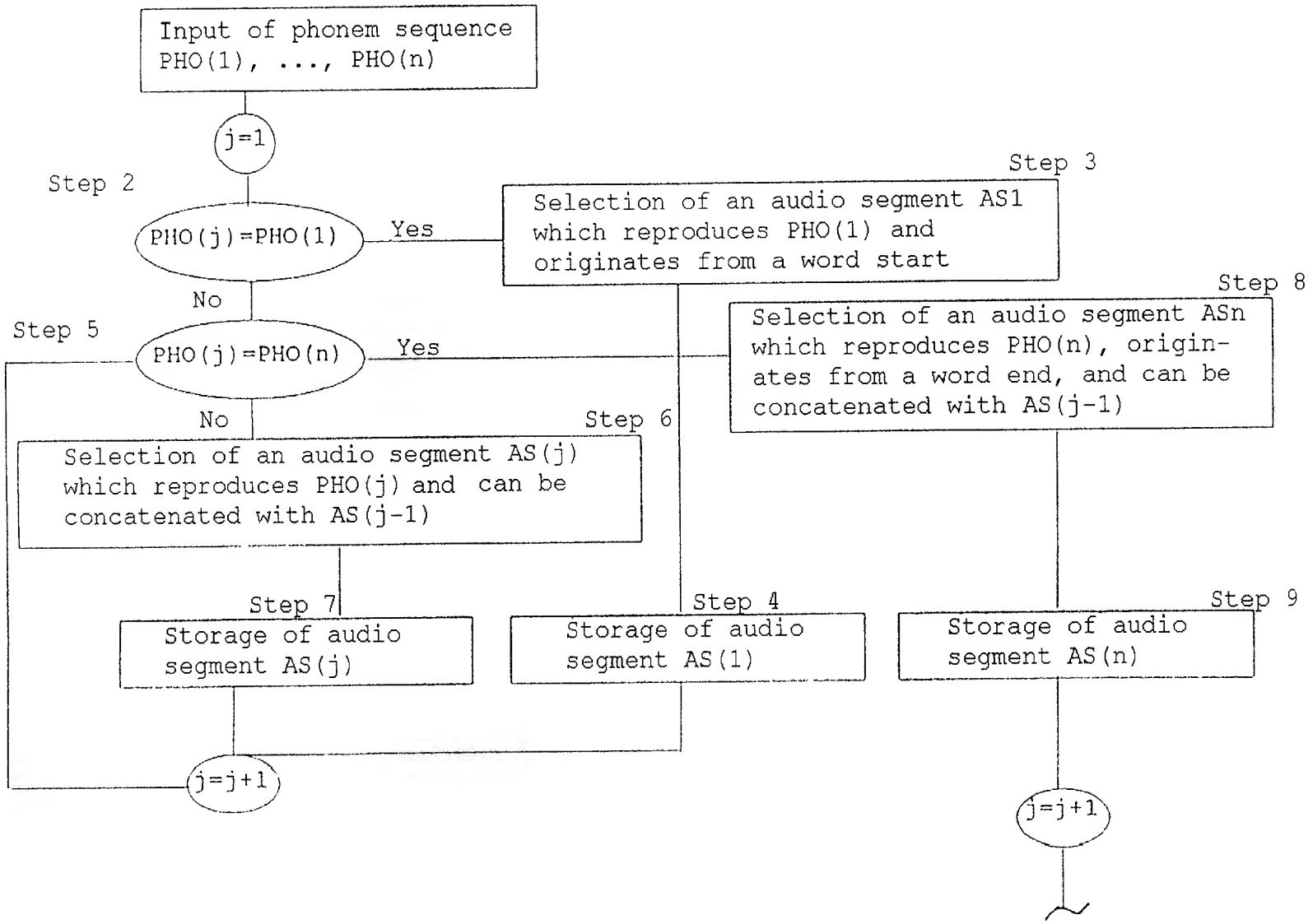
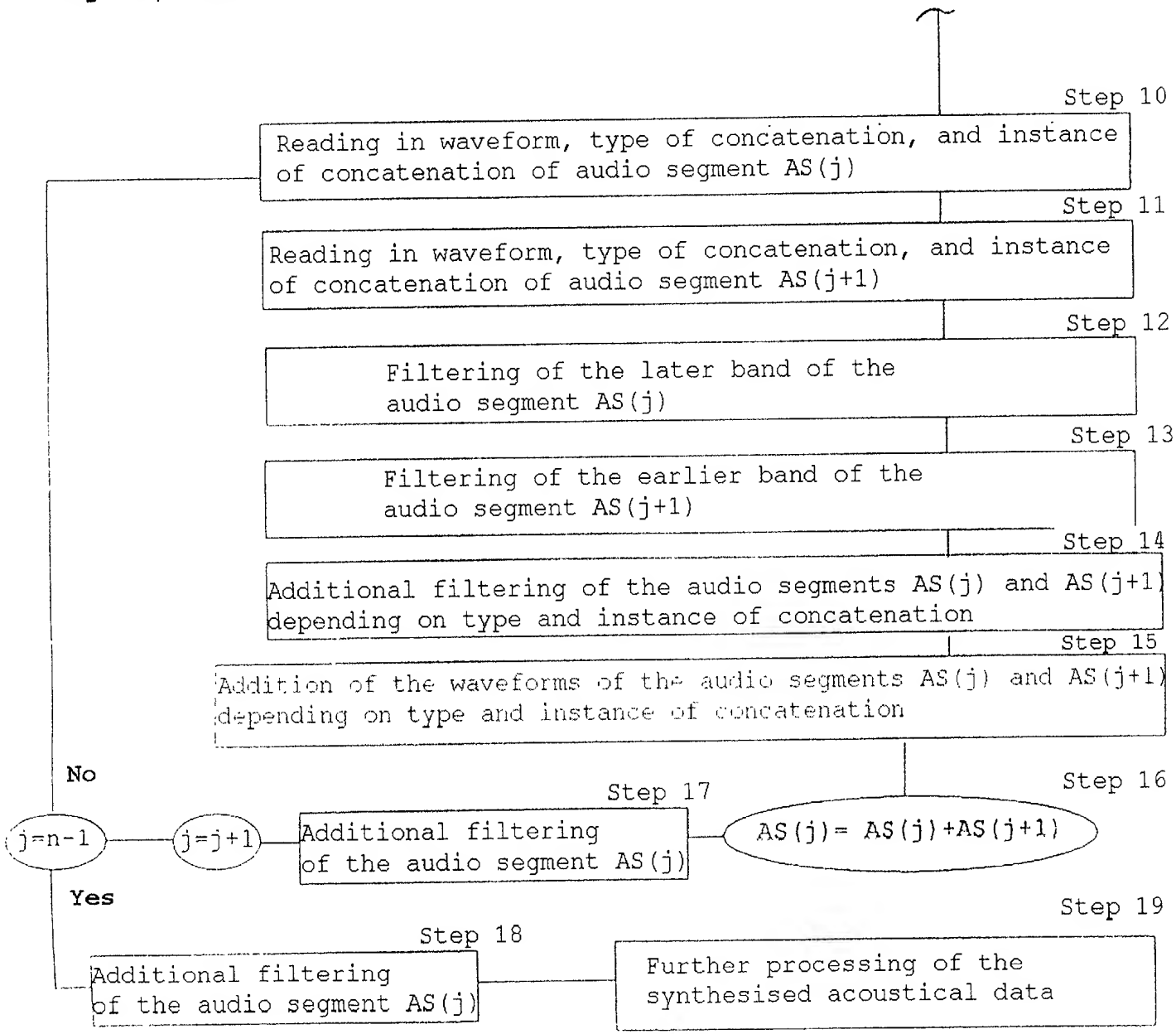


Fig. 4, Part 2



Docket No
87977.029101

Declaration and Power of Attorney For Patent Application

English Language Declaration

As a below named inventor, I hereby declare that:

My residence, post office address and citizenship are as stated below next to my name,

I believe I am the original, first and sole inventor (if only one name is listed below) or an original, first and joint inventor (if plural names are listed below) of the subject matter which is claimed and for which a patent is sought on the invention entitled
METHOD AND DEVICE FOR THE CONCATENATION OF AUDIOSEGMENTS, TAKING INTO ACCOUNT COARTICULATION

the specification of which

(check one)

☒ is attached hereto

☐ was filed on _____ as United States Application No. or PCT International Application Number _____

and was amended on _____
(if applicable)

I hereby state that I have reviewed and understand the contents of the above identified specification, including the claims, as amended by any amendment referred to above.

I acknowledge the duty to disclose to the United States Patent and Trademark Office all information known to me to be material to patentability as defined in Title 37, Code of Federal Regulations, Section 1.56.

I hereby claim foreign priority benefits under Title 35, United States Code, Section 119(a)-(d) or Section 365(b) of any foreign application(s) for patent or inventor's certificate, or Section 365(a) of any PCT International application which designated at least one country other than the United States, listed below and have also identified below, by checking the box, any foreign application for patent or inventor's certificate or PCT International application having a filing date before that of the application on which priority is claimed.

Prior Foreign Application(s)			Priority Not Claimed
198 37 661.8	DE	August 19, 1998	☐
(Number)	(Country)	(Day/Month/Year Filed)	
			☐
(Number)	(Country)	(Day/Month/Year Filed)	
			☐
(Number)	(Country)	(Day/Month/Year Filed)	

I hereby claim the benefit under 35 U.S.C Section 119(e) of any United States provisional application(s) listed below:

None	
(Application Serial No.)	(Filing Date)
(Application Serial No.)	(Filing Date)
(Application Serial No.)	(Filing Date)

I hereby claim the benefit under 35 U. S. C. Section 120 of any United States application(s), or Section 365(c) of any PCT International application designating the United States, listed below and, insofar as the subject matter of each of the claims of this application is not disclosed in the prior United States or PCT International application in the manner provided by the first paragraph of 35 U.S.C. Section 112, I acknowledge the duty to disclose to the United States Patent and Trademark Office all information known to me to be material to patentability as defined in Title 37, C. F. R., Section 1.56 which became available between the filing date of the prior application and the national or PCT International filing date of this application:

PCT/EP99/06081	August 19, 1999	Pending
(Application Serial No.)	(Filing Date)	(Status) (patented, pending, abandoned)
(Application Serial No)	(Filing Date)	(Status) (patented, pending, abandoned)
(Application Serial No.)	(Filing Date)	(Status) (patented, pending, abandoned)

I hereby declare that all statements made herein of my own knowledge are true and that all statements made on information and belief are believed to be true; and further that these statements were made with the knowledge that willful false statements and the like so made are punishable by fine or imprisonment, or both, under Section 1001 of Title 18 of the United States Code and that such willful false statements may jeopardize the validity of the application or any patent issued thereon.

POWER OF ATTORNEY: As a named inventor, I hereby appoint the following attorney(s) and/or agent(s) to prosecute this application and transact all business in the Patent and Trademark Office connected therewith. *(list name and registration number)*

6

Thomas R. FitzGerald	Reg. No. 26,730
Ronald S. Kareken	Reg. No. 20,573
Lee J. Fleckenstein	Reg. No. 36,136
Laurence S. Roach	Reg. No. 45,044
Stephen J. Sand	Reg. No. 34,716
Ronald J. Kisicki	Reg. No. 38,205

Send Correspondence to: Thomas R. FitzGerald, Esq.
Jaeckle Fleischmann & Mugel, LLP
39 State Street
Rochester, New York 14614-1310

Direct Telephone Calls to: *(name and telephone number)*
Thomas R. FitzGerald, Esq. Tel (716) 262-3640 - Fax (716) 262-4133

1

Full name of sole or first inventor
Christoph Buskies

Sole or first inventor's signature — Date
Christoph Buskies — 10.4.01

Residence
~~Alsenstrasse 21, 22769 Hamburg, Germany~~ Halenkamp 6, 22765 Hamburg, Germany

Citizenship
German DEX

Post Office Address
Same as Above

Full name of second inventor, if any

Second inventor's signature — Date

Residence

Citizenship

Post Office Address